(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 October 2003 (23.10.2003)

PCT

(10) International Publication Number
WO 03/087766 A2

(51) International Patent Classification⁷: G01N

(21) International Application Number: PCT/US03/10783

(22) International Filing Date: 4 April 2003 (04.04.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/370,895 5 April 2002 (05.04.2002) US

(71) Applicant (*for all designated States except US*): **THE GOVERNMENT OF THE UNITED STATES OF AMERICA, as represented by THE SECRETARY OF THE DEPARTMENT OF HEALTH AND HUMAN SERVICES** [US/US]; 6011 Executive Boulevard, Suite 325, Rockville, MD 20852 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): **WANG, Xin, Wei** [US/US]; 11409 Crownwood Lane, Rockville, MD 20850 (US). **YE, Qing-Hai** [CN/CN]; Xi Ying Road 33-22-22, Apt. 301, Pu Dong New Area, 200126 Shanghai (CN). **KIM, Jin, Woo** [KR/US]; 12030 Chase Crossing Cir., #404, Rockville, MD 20852 (US).

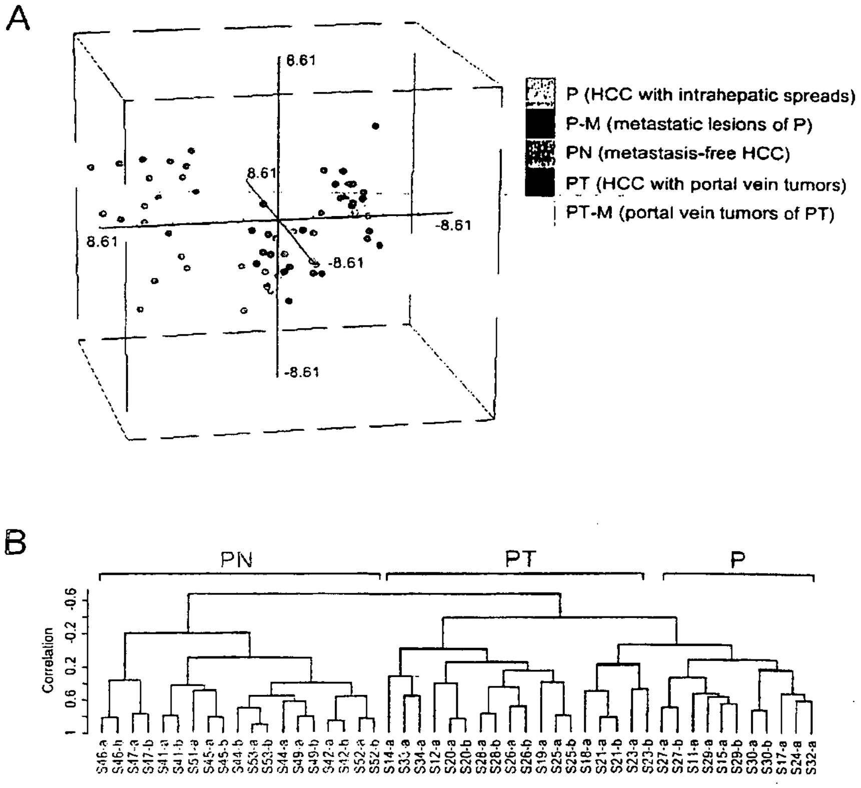
(74) Agents: **WEBER, Kenneth, A.** et al.; TOWNSEND AND TOWNSEND AND CREW LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111-3834 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),

[Continued on next page]

(54) Title: METHODS OF DIAGNOSING POTENTIAL FOR METASTASIS OR DEVELOPING HEPATOCELLULAR CARCINOMA AND OF IDENTIFYING THERAPEUTIC TARGETS



(57) Abstract: The present invention relates to methods for diagnosing the metastatic potential of hepatocellular carcinoma (HCC) in HCC patients and methods for diagnosing the potential of developing HCC in patients with chronic liver diseases. A computer readable medium, a digital computer, and a system useful for such diagnosis are also provided. Further disclosed are methods for identifying potential therapeutic targets for treating metastasis in HCC patients and methods for preventing HCC in patients with chronic liver diseases. In addition, the invention provides methods for inhibiting metastasis in HCC patients by suppressing the function of one therapeutic target, osteopontin, and methods for preventing the development of HCC in patients with chronic liver diseases by suppressing the function of one therapeutic target, EpCAM. Pharmaceutical compositions containing agents capable of inhibiting the functions of osteopontin or EpCAM are also disclosed.

WO 03/087766 A2

Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Methods of Diagnosing Potential for Metastasis or Developing Hepatocellular Carcinoma and of Identifying Therapeutic Targets

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application claims the benefit of priority to U.S. Provisional Patent Application No. 60/370,895, filed April 5, 2002, the entire contents of which are hereby incorporated by reference.

STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0002] This invention is owned by the United States of America as represented by the Secretary of Health and Human Services.

BACKGROUND OF THE INVENTION

[0003] Hepatocellular carcinoma (HCC) is one of the most common and aggressive malignancies worldwide with a curable rate of less than 5%. The high mortality is mainly due to the occurrence of intra-hepatic metastases. Little is known about the molecular basis of intra-hepatic metastasis or about specific therapeutic targets in these patients.

[0004] Within the past decade, several technologies have made it possible to monitor the expression level of a large number of transcripts at any one time (*see, e.g.*, Schena et al., *Science* **270**:467-470, 1995; Lockhart et al., *Nature Biotechnology* **14**:1675-1680, 1996; Blanchard et al., *Nature Biotechnology* **14**:1649, 1996; and U.S. Pat. No. 5,569,588). In organisms for which the complete genome is known, it is possible to analyze the transcripts of all genes within the cell. With other organisms, such as human, for which there is an increasing knowledge of the genome, it is possible to simultaneously monitor large numbers of the genes within the cell. Such monitoring technologies have been applied to the identification of genes which are up regulated or down regulated in various diseased or physiological states, the analyses of members of signaling cellular states, and the identification of targets for various drugs.

[0005] The present inventors analyzed the expression of 9,180 genes in HCC tissues from 40 patients without or with accompanying intra-hepatic metastases. Using a supervised machine learning algorithm to classify patients based on their gene expression signatures, a molecular signature has been generated for the first time that correctly classifies patients with or without metastases and have identifies genes that are mostly relevant to the prediction of outcome including patient survival. The gene expression signature of primary HCCs with accompanying metastasis is very similar to that of their corresponding metastases, suggesting that the genes favoring metastasis progression likely have been initiated in the primary tumors. Moreover, osteopontin (OPN) is overexpressed in primary HCC with intra-hepatic metastasis and a neutralizing antibody against osteopontin is shown to block invasion of highly metastatic HCC cells in an *in vitro* assay of invasion. These data identify osteopontin both as a diagnostic marker and a therapeutic target for metastatic HCC.

[0006] The expression of 9,180 genes has also been analyzed in tumor samples from 54 HCC patients and in 59 non-cancerous liver samples from patients with severe liver diseases and at high risk for developing HCC or at low risk for developing HCC. The high risk group includes patients diagnosed with hepatitis B, hepatitis C, hemochromatosis, and Wilson's disease. The low risk group includes patients diagnosed with alcoholic liver disease, autoimmune hepatitis, and primary biliary cirrhosis. A comparison of the gene expression levels between the high risk and low risk groups has identified a set of significant genes that would differentiate between the high risk and low risk groups. Filtering the set of significant genes using expression data from HCC samples has identified subsets of genes enriched with HCC-related molecular signatures and useful for classifying samples. In addition, EpCAM is among the most significant genes whose overexpression positively correlates to the risk of developing HCC in a patient with a severe liver disease and the inhibition of its expression has been shown to lead to growth suppression in HCC cells. Thus, EpCAM has been identified as a diagnostic marker for predicting the risk of developing HCC as well as a therapeutic target for preventing the onset of HCC in patients suffering from chronic liver diseases.

BRIEF SUMMARY OF THE INVENTION

[0007] One aspect of the present invention relates to a method for identifying potential therapeutic targets for inhibiting metastasis in a patient suffering from HCC or for preventing the development of HCC in a patient suffering from a chronic liver disease.

[0008] The method for identifying potential therapeutic targets for inhibiting metastasis in an HCC patient includes the steps of: a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a metastatic HCC patient; b) capturing markers from the sample and generating a first signal; c) repeating steps a) and b) with a sample from a non-metastatic HCC patient and thereby generating a second signal; and d) comparing the first and second signals and thereby identifying a subset of cellular markers whose level is different in the first and second signals, wherein the subset of cellular markers are potential therapeutic targets for treating HCC metastasis in an HCC patient. In some embodiments, a signal generated from a normal non-cancerous sample on an array identical to the array of step a) is subtracted in steps b) and c) to generate the first and second signals.

[0009] The method for identifying potential therapeutic targets for preventing the onset of HCC in a patient with a chronic liver disease includes the steps of: a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a patient with a chronic liver disease and a high risk of developing HCC; b) capturing markers from the sample and generating a first signal; c) repeating steps a) and b) with a sample from a patient with a chronic liver disease and a low risk of developing HCC and thereby generating a second signal; and d) comparing the first and second signals and thereby identifying a subset of cellular markers whose level is different in the first and second signals, wherein the subset of cellular markers are potential therapeutic target for preventing HCC in a patient with a chronic liver disease. In some embodiments, a signal generated from a normal non-cancerous sample on an array identical to the array of step a) is subtracted in steps b) and c) to generate the first and second signals.

[0010] Another aspect of the present invention relates to a method for predicting the metastatic potential in an HCC patient or for predicting the risk of developing HCC in a patient with a chronic liver disease.

[0011] The method for predicting the metastatic potential in an HCC patient includes the steps of: a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a metastatic HCC patient, the set of cellular markers comprising at least ten genes or proteins encoded by genes independently selected from the genes of Table 2; b) capturing markers from the sample; c) generating a first signal from the captured markers of step b); d) repeating steps a) to c) with a sample from a non-metastatic HCC patient and thereby generating a second signal; e) repeating steps a) to c) with a sample from an HCC patient with unknown metastatic potential and thereby generating a third signal; and f) comparing the third signal to the first and the second signals and thereby determining the metastatic potential of the HCC patient of step e). In some embodiments, the set of cellular markers includes at least 20, preferably 50, more preferably 100, and most preferably all genes or proteins encoded by genes independently selected from the genes of Table 2. In other embodiments, the set of cellular markers includes the genes or proteins encoded by genes of Table 4 or Unigene numbers Hs.313, Hs.69707, Hs.222, Hs.63984, Hs.75573, Hs.177687, Hs.69707, Hs.222, Hs.323712, and Hs.63984. Preferably, the sample of steps a) and b), the sample of step d), and the sample of step e) are liver tissue extracts. In a preferred embodiment, the array of step a) is a genomic array. In another preferred embodiment, the array of step a) is a proteomic array.

[0012] The method for predicting the risk of developing HCC in a patient suffering from a chronic liver disease includes the steps of: a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a patient with a chronic liver disease and a high risk of HCC, the set of cellular markers comprising at least ten genes or proteins encoded by genes independently selected from the genes of Table 5; b) capturing markers from the sample; c) generating a first signal from the captured markers of step b); d) repeating steps a) to c) with a sample from a patient with a chronic liver disease and a low risk of HCC and thereby generating a second signal; e) repeating steps a) to c) with a sample from a patient with a chronic liver disease and an unknown risk of HCC and thereby generating a third signal; and f) comparing the third signal to the first and the second signals and thereby determining the risk of developing HCC in the patient of step e). In some embodiments, the set of cellular markers comprises at least 20, preferably 50, more preferably 100, and most preferably all genes or proteins encoded by genes independently selected from the genes of Table 5. In some othe embodiments, the set of cellular markers comprises the genes or proteins encodec by genes of Table 6 or Table 7. Preferably, the sample of steps a) and b), the sample of step d), and the sample of step e) are liver tissue extracts. In one preferred embodiment, the array of step a) is a genomic array. In another preferred embodiment, the array of step a) is a proteomic array. In some embodiments, the patient with a high risk of developing HCC suffers from hepatitis B infection, hepatitis C, hemachromatosis, or Wilson's disease. In other embodiments, the patient with a low risk of HCC suffers from alcoholic liver disease, autoimmune hepatitis, or primary biliary cirrhosis. In yet other embodiments, the patient whose risk of developing HCC is being assessed suffers from hepatitis B, hepatitis C, hemochromatosis, Wilson's disease, alcoholic liver disease, autoimmune hepatitis, or primary biliary cirrhosis.

[0013] Yet another aspect of the invention relates to a method for inhibiting metastasis in an HCC patient as well as a method for inhibiting the development of HCC in a patient with a chronic liver disease. The method for inhibiting HCC metastasis in an HCC patient includes the step of suppressing OPN activity. In some embodiments, suppression of OPN activity is accomplished by inhibiting OPN expression, preferably using an antisense polynucleotide specific for OPN. In other embodiments, suppression of OPN activity is accomplished by inhibiting the specific binding between OPN and OPN receptor, preferably using an anti-OPN antibody. The method for preventing the onset of HCC in a patient with a chronic liver disease includes the step of suppressing EpCAM activity. In some embodiments, suppression of EpCAM activity is accomplished by inhibiting EpCAM expression, preferably using an antisense polynucleotide or a small inhibitory RNA molecule specific for EpCAM. In other embodiments, suppression of EpCAM activity is accomplished by inhibiting the specific binding between EpCAM and EpCAM receptor, preferably using an anti-EpCAM antibody.

[0014] A still further aspect of the present invention relates to a computer readable medium, a digital computer, and a system for accessing the metastatic potential in an HCC patient or the risk of developing HCC in a patient with a chronic liver disease.

[0015] The computer readable medium for assessing the metastatic potential in an HCC patient includes: a) code for a first data set, derived from a first signal from an array comprising capture reagents for a set of cellular markers after contact with a sample from a metastatic HCC patient, the set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 2; b) code for a second data set, derived from a second signal from an array identical to the array of a) after contact with a sample from a non-metastatic HCC patient; c) code for a third data set, derived from a third signal from an array identical to the array of a) after contact with a sample from a HCC patient with unknown metastatic potential; and d) code for comparing the third data set with the first and second data sets. A digital computer containing the claimed computer readable medium for assessing HCC metastatic potential in an HCC patient is also provided. Further provided is a system containing such a digital computer, a chip with an array comprising capture reagents for a set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 2, and a reader capable of registering a signal from the array after contact with a sample.

[0016] The computer readable medium for assessing the risk of developing HCC in a patient with a chronic liver disease includes: a) code for a first data set, derived from a first signal from an array comprising capture reagents for a set of cellular markers after contact with a sample from a patient with a chronic liver disease and a high risk of HCC, the set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 5; b) code for a second data set, derived from a second signal from an array identical to the array of a) after contact with a sample from a patient with a chronic liver disease and a low risk of HCC; c) code for a third data set, derived from a third signal from an array identical to the array of a) after contact with a sample from a patient with a chronic liver disease and an unknown risk of HCC; and d) code for comparing the third data set with the first and second data sets. A digital computer containing the claimed computer readable medium for assessing the risk of develop HCC in a patient with a chronic liver disease is also provided. Further provided is a system containing such a digital computer, a chip with an array comprising capture reagents for a set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 5, and a reader capable of registering a signal from the array after contact with a sample.

DEFINITIONS

[0017] Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by a person skilled in the art to which this invention belongs. The following references provide one of skill with a general definition of many of the terms used in this invention: Singleton *et al.*, *Dictionary of Microbiology and Molecular Biology* (2nd ed. 1994); *The Cambridge Dictionary of Science and Technology* (Walker ed., 1988); *The Glossary of Genetics*, 5th Ed., R. Rieger *et al.* (eds.), Springer Verlag (1991); and Hale & Marham, *The Harper Collins Dictionary of Biology* (1991). As used herein, the following terms have the meanings ascribed to them unless specified otherwise.

[0018] The term "hepatocellular carcinoma" or "HCC" as used herein refer to the major type of carcinoma of the liver that accounts for more than 90% of all primary liver cancers. Hepatocellular carcinomas range from well differentiated to highly anaplastic undifferentiated lesions. Hepatocellular carcinomas may exist as single intra-hepatic lesions (non-metastatic), multifocal intra-hepatic metastasis or as extra-hepatic metastasis.

[0019] "High risk precancerous diseases" refer to a group of epidemiologically defined diseases that are associated with a high probability of developing HCC. These diseases include chronic hepatitis B infection, hepatitis C infection, hemochromatosis, and Wilson's disease.

[0020] "Low risk precancerous diseases" refer to a group of epidemiologically defined diseases, that are associated with a low risk of developing HCC. These diseases include alcoholic liver disease, autoimmune hepatitis, and primary biliary cirrhosis.

[0021] The term "metastasis" or "metastatic" refers to the ability of a cancer cell to invade surrounding tissues, to enter the circulatory system and to establish malignant growths at new sites.

[0022] "Non-Metastatic" refers to tumors that do not spread beyond their original site of development and specifically do not enter the circulatory system and establish malignant growths at new sites.

[0023] The term "non-cancerous" refers to a biological sample or tissue sample in which the cells in the sample exhibit a normal or non-pathological phenotype when analyzed visually, by microscope, immunohistologically, immunologically, or molecularly using antibody or nucleic acid probes designed to detect pathological conditions.

[0024] The term "normal" refers to a biological sample or tissue sample in which the sample is obtained from an individual who has not been diagnosed with HCC or high risk, or low risk precancerous diseases.

[0025] The term "capture reagent" refers to any type of moiety that binds to a specific nucleic acid or protein marker. Typically the binding of the marker to the capture reagent can be controlled by the conditions used during the binding process. For example, the binding of a nucleic acid marker to a cognate oligonucleotide is controlled by the hybridization conditions used. Stringent hybridizations conditions will only allow a nucleic acid marker that has high homology *e.g.* 95%-100% identity with the oligonucleotide to bind to the oligonucleotide.

[0026] "Array" refers to a plurality of capture reagents bound to a substrate, *e.g.*, a solid support, which will bind to their cognate markers. For example, the array may be composed of nucleic acid molecules, protein molecules or any other reagent that will specifically bind a nucleic acid, protein or polypeptide isolated from a biological sample. The capture reagents are preferentially bound in an addressable fashion such that when the cognate marker is bound to the capture reagent, the amount of binding may be quantified.

[0027] "DNA microarray" refers to an array in which the capture reagents are nucleic acid molecules. Typically, a DNA microarray is composed of DNA oligonucleotides of a defined length which can hybridize to DNA, cDNA or RNA molecules under defined conditions. DNA oligonucleotides may be short pieces of nucleic acid ranging is size from 15-50 bases or they may be longer pieces of nucleic acids ranging in size from 500-1000 bases or longer. DNA microarrays may be composed of hundreds or thousands of different nucleic acid molecules each of which is located on the array in a defined position. Binding of the marker to the DNA microarray is usually quantified when the marker is labeled with a detectable moiety. The term DNA microarray is used interchangeably with the term "genomic array"

[0028] "Protein array" refers to an array in which the capture reagents will bind protein markers. Typically these reagents may be polyclonal or monoclonal antibodies that bind specific proteins. Alternatively, any protein, peptide, nucleic acid or other molecule or surface which will specifically bind to a protein may be used in a protein array. These arrays usually contain hundreds or thousands of different capture reagents in addressable locations. Binding of the markers to the capture reagent on the protein array is usually quantified when the marker is labeled with a detectable moiety. The term protein array is used interchangeably with "proteomic array".

[0029] "Gene expression profile" refers to the all of the genes that are expressed in a tissue sample compared to a reference sample. The level of gene expression of genes in a gene expression profile is determined by comparing the level of expression in a test sample *e.g.* an HCC tumor sample or a sample obtained from a patient diagnosed with severe liver disease to the level of expression in a reference sample. The reference sample used for determining the metastatic potential of an HCC tumor is non-cancerous liver tissue or liver tissue obtained from a patient who has not been diagnosed with HCC. The reference sample used for determining the potential for developing HCC in patients diagnosed with severe liver disease is liver tissue obtained from patients who have not been diagnosed with severe liver disease. Genes in the test sample may be over expressed or under expressed relative to the reference sample.

[0030] "Metastatic gene expression predictor" refers to the expression of a specific cluster of genes correlated with the diagnosis of metastatic HCC. The metastatic gene expression predictor is generated by comparing the gene expression profile of a test sample obtained from a non-metastatic HCC sample to the gene expression profile obtained from a metastatic HCC sample followed by a cluster and classification analysis using a defined algorithm or set of algorithms. The number of genes present may vary depending on the clustering algorithm used or depending on a parameter in the algorithm *e.g.* p-level = 0.001 vs. 0.022.

[0031] "HCC gene expression predictor" refers to the expression of a specific cluster of genes correlated with the diagnosis of patients likely to develop HCC. The HCC gene expression predictor is generated by comparing the gene expression profile of a test sample obtained from a non-metastatic liver sample obtained from a patient with a high risk for developing HCC to the gene expression profile obtained from a non-metastatic liver sample obtained from a patient having a low risk of developing HCC followed by a cluster and classification analysis using a defined algorithm or set of algorithms. The number of genes present may vary depending on the clustering algorithm used or depending on a parameter in the algorithm *e.g.* p-level = 0.001 vs. 0.022.

[0032] "UG Cluster" used in Tables 2-7 refers to the UniGene data base compiled by the National Center for Biological Information ("NCBI"). Each accession number in the UniGene data base is a compilation of all of the nucleotide and amino acid sequence data available for a specific nucleotide sequence. For example, each UG Cluster accession number may provide links to GeneBank or other data base which in turn provide nucleotide sequences encoding a partial or full length cDNA for a gene. Alternatively the links may provide genomic or EST sequence data or amino acid sequence information. Each UG Cluster accession number provides unique sequence information for the specific gene, nucleic acid or amino acid sequence identified.

[0033] "Ostoepontin" refers to a secreted phosphoprotein encoded by SEQ ID NO:1 or a conservative variant thereof, which may also be found in Genbank accession number NM_000582. Nucleic acid and amino acid sequence information may also be found in the National Center for Biological Information ("NCBI") UniGene data base under accession number Hs.313 at NCBI web site. This site lists 9 mRNA/genomic DNA sequences and over 900 expressed sequence tags. Osteopontin is an extracellular protein associated with the bone matrix and associated with atherosclerotic plaques. Full length osteopontin protein contains an RGD amino acid sequence that functions as an integrin binding site. Osteopontin is a major ligand for the vitronectin receptor. "OPN" is used interchangeably with osteopontin and refers either to the protein, the gene encoding the protein or fragments thereof.

[0034] "EpCAM" is a 40 kDa glycoprotein that functions as an Epithelial Cell Adhesion Molecule. It is also identified as tumor-associated calcium signal transducer or TACSTD1, with a Unigene Cluster number of Hs.692. EpCAM is encoded by the *GA733-2* gene, which is located on human chromosome 4q. A transmembrane protein expressed in cells of epithelial origin, EpCAM mediates Ca^{2+}-independent homotypic cell-cell adhesion and is specifically recognized by a number of well known monoclonal antibodies (mAb), such as 17-1A, 323/A3, KS1/4, GA733, MOC31, etc.

[0035] The term "Marker" in the context of the present invention refers to a nucleic acid sequence or a gene encoding a polypeptide (of a particular apparent molecular weight) which is differentially present in a sample taken from patients having metastatic HCC or a predisposition for HCC as compared to a comparable sample taken from control subjects (*e.g.*, a person with non-metastatic HCC or a negative diagnosis or undetectable cancer, normal or healthy subject). Marker may also refer to a polypeptide or protein encoded by a nucleic acid sequence or gene which is differentially present in a sample taken from patients having metastatic HCC or a predisposition for HCC as compared to a comparable sample taken from control subjects (*e.g.*, a person with non-metastatic HCC or a negative diagnosis or undetectable cancer, normal or healthy subject). Markers of the present invention include the genes and their encoded proteins identified by UG Cluster number in Tables 2-7 infra.

[0036] The term "sample" as used herein is a sample of biological tissue or fluid that will be used to determine a gene expression profile, a source of markers, or that contains a protein of interest (such as osteopontin or EpCAM) or a nucleic acid encoding such protein. Such samples include, but are not limited to, various types of tissue isolated from humans, and may also include sections of tissues such as frozen sections or paraffin sections taken for histological purposes. Tissues include liver samples and fluid samples include blood, serum, plasma, urine, and other bodily fluids. A preferred sample used for practicing the present invention is a lysate of cells extracted from a tissue of interest, *e.g.*, liver. Such a cell lysate may be prepared using a variety of methods known to those skilled in the art, depending on the form in which a cellular marker is to be detected and examined, *e.g.*, as a nucleic acid such as mRNA, as a protein, or as a molecule with other measurable biological characteristics such as an enzymatic activity.

[0037] The phrase "functional effects" in the context of assays for testing compounds that regulate the biological activity of a protein of interest, *e.g.*, osteopontin or EpCAM, includes the determination of any parameter that is directly or indirectly related to or under the influence of OPN or EpCAM, such as the level of mRNA encoding the proteins, the level of the proteins, as well as their functional, physical, and chemical effects (*e.g.*, their ability to specifically interact with their naturally binding partners, such as other proteins, nucleic acids, or any other molecules, their ability to mediate signal transduction that may affect cellular events such as cell proliferation, differentiation, apoptosis, secretion, adhesion, and the like).

[0038] "Nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid, and which are metabolized in a manner similar to the reference nucleotides. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, peptide-nucleic acids (PNAs). The term encompasses nucleic acids isolated from biological samples and synthetic oligonucleotides.

[0039] Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof *(e.g.*, degenerate codon substitutions) and complementary sequences, as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al., Nucleic Acid Res.* 19:5081, 1991; Ohtsuka *et al., J. Biol. Chem.* 260:2605-2608, 1985; Rossolini *et al., Mol. Cell. Probes* 8:91-98, 1994). The term nucleic acid is used interchangeably with gene, cDNA, mRNA, oligonucleotide, and polynucleotide.

[0040] The terms "polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymer.

[0041] The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, *e.g.*, hydroxyproline, γ-carboxyglutamate, and O-phosphoserine. Amino acid analogs refer to compounds that have the same basic chemical structure as a naturally occurring amino acid, *i.e.*, an α carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, *e.g.*, homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs have modified R groups (*e.g.*, norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refer to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid.

[0042] Amino acids may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

[0043] "Conservatively modified variants" applies to both amino acid and nucleic acid sequences. With respect to particular nucleic acid sequences, conservatively modified variants refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given protein. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation of the nucleic acid. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine, and TGG, which is ordinarily the only codon for tryptophan) can be modified to yield a functionally identical molecule. Accordingly, each silent variation of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

[0044] As to amino acid sequences, one of skill will recognize that individual substitutions, deletions or additions to a nucleic acid, peptide, polypeptide, or protein sequence which alters, adds or deletes a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration results in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. Such conservatively modified variants are in addition to and do not exclude polymorphic variants, interspecies homologs, and alleles of the invention.

[0045] The following eight groups each contain amino acids that are conservative substitutions for one another:

1) Alanine (A), Glycine (G);
2) Aspartic acid (D), Glutamic acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V);
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W);
7) Serine (S), Threonine (T); and
8) Cysteine (C), Methionine (M)

(*see, e.g.*, Creighton, *Proteins*, 1984).

[0046] Macromolecular structures such as polypeptide structures can be described in terms of various levels of organization. For a general discussion of this organization, *see, e.g.*, Alberts *et al.*, *Molecular Biology of the Cell* (3rd ed., 1994) and Cantor and Schimmel, *Biophysical Chemistry Part I: The Conformation of Biological Macromolecules* (1980). "Primary structure" refers to the amino acid sequence of a particular peptide. "Secondary structure" refers to locally ordered, three dimensional structures within a polypeptide. These structures are commonly known as domains. Domains are portions of a polypeptide that form a compact unit of the polypeptide and are typically 50 to 350 amino acids long. Typical domains are made up of sections of lesser organization such as stretches of β-sheet and α-helices. "Tertiary structure" refers to the complete three dimensional structure of a polypeptide monomer. "Quaternary structure" refers to the three dimensional structure formed by the noncovalent association of independent tertiary units. Anisotropic terms are also known as energy terms.

[0047] "Antibody" refers to a polypeptide comprising a framework region from an immunoglobulin gene or fragments thereof that specifically binds and recognizes an antigen. The recognized immunoglobulin genes include the kappa, lambda, alpha, gamma, delta, epsilon, and mu constant region genes, as well as the myriad immunoglobulin variable region genes. Light chains are classified as either kappa or lambda. Heavy chains are classified as gamma, mu, alpha, delta, or epsilon, which in turn define the immunoglobulin classes, IgG, IgM, IgA, IgD and IgE, respectively.

[0048] An exemplary immunoglobulin (antibody) structural unit comprises a tetramer. Each tetramer is composed of two identical pairs of polypeptide chains, each pair having one "light" (about 25 kDa) and one "heavy" chain (about 50-70 kDa). The N-terminus of each chain defines a variable region of about 100 to 110 or more amino acids primarily responsible for antigen recognition. The terms variable light chain (V_L) and variable heavy chain (V_H) refer to these light and heavy chains respectively.

[0049] Antibodies exist, *e.g.*, as intact immunoglobulins or as a number of well-characterized fragments produced by digestion with various peptidases. Thus, for example, pepsin digests an antibody below the disulfide linkages in the hinge region to produce F(ab)'$_2$, a dimer of Fab which itself is a light chain joined to V_H-C_H1 by a disulfide bond. The F(ab)'$_2$ may be reduced under mild conditions to break the disulfide linkage in the hinge region, thereby converting the F(ab)'$_2$ dimer into an Fab' monomer. The Fab' monomer is essentially Fab with part of the hinge region (*see Fundamental Immunology* (Paul ed., 3d ed. 1993). While various antibody fragments are defined in terms of the digestion of an intact antibody, one of skill will appreciate that such fragments may be synthesized *de novo* either chemically or by using recombinant DNA methodology. Thus, the term antibody, as used herein, also includes antibody fragments either produced by the modification of whole antibodies, or those synthesized *de novo* using recombinant DNA methodologies (*e.g.*, single chain Fv) or those identified using phage display libraries (*see*, *e.g.*, McCafferty *et al.*, *Nature* **348**:552-554, 1990).

[0050] For preparation of monoclonal or polyclonal antibodies, any technique known in the art can be used (*see*, *e.g.*, Kohler & Milstein, *Nature* 256:495-497 (1975); Kozbor *et al.*, *Immunology Today* 4: 72 (1983); Cole *et al.*, pp. 77-96 in *Monoclonal Antibodies and Cancer*

Therapy (1985)). Techniques for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce antibodies to polypeptides of this invention. Also, transgenic mice, or other organisms such as other mammals, may be used to express humanized antibodies. Alternatively, phage display technology can be used to identify antibodies and heteromeric Fab fragments that specifically bind to selected antigens (*see, e.g.*, McCafferty *et al.*, *supra*; Marks *et al.*, *Biotechnology* **10**:779-783, 1992).

[0051] A "chimeric antibody" is an antibody molecule in which (a) the constant region, or a portion thereof, is altered, replaced or exchanged so that the antigen binding site (variable region) is linked to a constant region of a different or altered class, effector function and/or species, or an entirely different molecule which confers new properties to the chimeric antibody, *e.g.*, an enzyme, toxin, hormone, growth factor, drug, etc.; or (b) the variable region, or a portion thereof, is altered, replaced or exchanged with a variable region having a different or altered antigen specificity.

[0052] An "anti-OPN antibody" is an antibody or antibody fragment that specifically binds a polypeptide encoded by the OPN gene, cDNA, or a subsequence thereof. An anti-EpCAM antibody is defined in a similar fashion.

[0053] A "receptor" as used herein encompasses any molecule that a particular protein, *e.g.*, OPN or EpCAM, can specifically bind and may thus include proteins, nucleic acids, carbohydrates, or any other molecules.

[0054] The term "immunoassay" is an assay that uses an antibody to specifically bind an antigen. The immunoassay is characterized by the use of specific binding properties of a particular antibody to isolate, target, and/or quantify the antigen.

[0055] The phrase "specifically (or selectively) binds" to an antibody or "specifically (or selectively) immunoreactive with," when referring to a protein or peptide, refers to a binding reaction that is determinative of the presence of the protein in a heterogeneous population of proteins and other biologics. Thus, under designated immunoassay conditions, the specified antibodies bind to a particular protein at least two times the background and do not substantially bind in a significant amount to other proteins present in the sample. Specific binding to an antibody under such conditions may require an antibody that is selected for its specificity for a particular protein. For example, polyclonal antibodies raised to OPN from specific species such as rat, murine, or human can be selected to obtain only those polyclonal antibodies that are specifically immunoreactive with OPN and not with other proteins, except for polymorphic variants and alleles of OPN. This selection may be achieved by subtracting out antibodies that cross-react with OPN molecules from other species. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays are routinely used to select antibodies specifically immunoreactive with a protein (*see, e.g.*, Harlow & Lane, *Antibodies, A Laboratory Manual*, 1988), for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity). Typically a specific or selective reaction will be at least twice background signal or noise and more typically more than 10 to 100 times background.

[0056] The phrase "differentially present" refers to differences in the quantity and/or the frequency of a marker present in a sample taken from a metastatic HCC tumor or liver samples of a patient at high risk for HCC as compared to a non-metastatic HCC sample or a liver sample from a patient at low risk for HCC respectively. For examples, a marker can be a polypeptide or nucleic acid which is present at an elevated level or at a decreased level in samples of metastatic HCC tumors or liver samples of someone at high risk for HCC compared to non-metastatic HCC samples or a liver sample from a patient at low risk for HCC respectively. Alternatively, a marker can be a polypeptide which is detected at a higher frequency or at a lower frequency in metastatic HCC tumors or liver samples of someone at high risk for HCC compared to non-metastatic HCC sample or a liver sample from a patient at low risk for HCC respectively. A marker can be differentially present in terms of quantity, frequency or both.

[0057] A polypeptide or nucleic acid is differentially present between the two samples if the amount of the polypeptide in one sample is statistically significantly different from the amount of the polypeptide in the other sample. For example, a polypeptide is differentially present between the two samples if it is present at least about 120%, at least about 130%, at least about 150%, at least about 180%, at least about 200%, at least about 300%, at least about 500%, at least about 700%, at least about 900%, or at least about 1000% greater than it is present in the other sample, or if it is detectable in one sample and not detectable in the other.

[0058] Alternatively or additionally, a polypeptide is differentially present between the two sets of samples if the frequency of detecting the polypeptide in the metastatic HCC tumors or liver samples of someone at high risk for HCC is statistically significantly higher or lower than in non-metastatic HCC samples or a liver sample from a patient at low risk for HCC respectively. For example, a polypeptide is differentially present between the two sets of samples if it is detected at least about 120%, at least about 130%, at least about 150%, at least about 180%, at least about 200%, at least about 300%, at least about 500%, at least about 700%, at least about 900%, or at least about 1000% more frequently or less frequently observed in one set of samples than the other set of samples.

[0059] "Diagnostic" means identifying the presence or nature of a pathologic condition or a predisposition for a pathologic condition such as HCC or HCC metastasis. Diagnostic methods differ in their sensitivity and specificity. The "sensitivity" of a diagnostic assay is the percentage of diseased individuals who test positive (percent of "true positives"). Diseased individuals not detected by the assay are "false negatives." Subjects who are not diseased and who test negative in the assay, are termed "true negatives." The "specificity" of a diagnostic assay is 1 minus the false positive rate, where the "false positive" rate is defined as the proportion of those without the disease who test positive. While a particular diagnostic method may not provide a definitive diagnosis of a condition, it suffices if the method provides a positive indication that aids in diagnosis.

[0060] A "test amount" of a marker refers to an amount of a marker present in a sample being tested. A test amount can be either in absolute amount (*e.g.*, μg/ml) or a relative amount (*e.g.*, relative intensity of signals).

[0061] A "diagnostic amount" of a marker refers to an amount of a marker in a subject's sample that is consistent with a diagnosis of metastatic HCC tumors or tissue samples of someone at high risk for HCC. A diagnostic amount can be either in absolute amount (*e.g.*, μg/ml) or a relative amount (*e.g.*, relative intensity of signals).

[0062] A "control amount" of a marker can be any amount or a range of amount which is to be compared against a test amount of a marker. For example, a control amount of a marker can be the amount of a marker in a person without metastatic HCC tumors or tissue samples of someone at low risk for HCC. A control amount can be either in absolute amount (*e.g.*, μg/ml) or a relative amount (*e.g.*, relative intensity of signals).

[0063] "Spectrometer probe" refers to a device that is removably insertable into a gas phase ion spectrometer and comprises a substrate having a surface for presenting a marker for detection. A spectrometer probe can comprise a single substrate or a plurality of substrates.

Terms such as ProteinChip®, ProteinChip® array, or chip are also used herein to refer to specific kinds of spectrometer probes.

[0064] "Substrate" or "probe substrate" refers to a solid phase onto which an adsorbent can be provided (*e.g.*, by attachment, deposition, *etc.*).

[0065] "Adsorbent" refers to any material capable of adsorbing a marker. The term "adsorbent" is used herein to refer both to a single material ("monoplex adsorbent") (*e.g.*, a compound or functional group) to which the marker is exposed, and to a plurality of different materials ("multiplex adsorbent") to which the marker is exposed. The adsorbent materials in a multiplex adsorbent are referred to as "adsorbent species." For example, an addressable location on a probe substrate can comprise a multiplex adsorbent characterized by many different adsorbent species (*e.g.*, anion exchange materials, metal chelators, or antibodies), having different binding characteristics. Substrate material itself can also contribute to adsorbing a marker and may be considered part of an "adsorbent."

[0066] "Adsorption" or "retention" refers to the detectable binding between an absorbent and a marker either before or after washing with an eluant (selectivity threshold modifier) or a washing solution.

[0067] "Eluant" or "washing solution" refers to an agent that can be used to mediate adsorption of a marker to an adsorbent. Eluants and washing solutions are also referred to as "selectivity threshold modifiers." Eluants and washing solutions can be used to wash and remove unbound materials from the probe substrate surface.

[0068] "Resolve," "resolution," or "resolution of marker" refers to the detection of at least one marker in a sample. Resolution includes the detection of a plurality of markers in a sample by separation and subsequent differential detection. Resolution does not require the complete separation of one or more markers from all other biomolecules in a mixture. Rather, any separation that allows the distinction between at least one marker and other biomolecules suffices.

[0069] "Gas phase ion spectrometer" refers to an apparatus that measures a parameter which can be translated into mass-to-charge ratios of ions formed when a sample is volatilized and ionized. Generally ions of interest bear a single charge, and mass-to-charge ratios are often simply referred to as mass. Gas phase ion spectrometers include, for example, mass spectrometers, ion mobility spectrometers, and total ion current measuring devices.

[0070] "Mass spectrometer" refers to a gas phase ion spectrometer that includes an inlet system, an ionization source, an ion optic assembly, a mass analyzer, and a detector.

[0071] "Laser desorption mass spectrometer" refers to a mass spectrometer which uses laser as means to desorb, volatilize, and ionize an analyte.

[0072] "Detect" refers to identifying the presence, absence, or amount of the object to be detected.

[0073] "Detectable moiety" or a "label" refers to a composition detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include ^{32}P, ^{35}S, fluorescent dyes, electron-dense reagents, enzymes (such as those commonly used in an ELISA, *e.g.*, horseradish peroxidase), biotin-streptavidin, digoxigenin, haptens and proteins for which antisera or monoclonal antibodies are available, or nucleic acid molecules with a sequence complementary to a target. The detectable moiety often generates a measurable signal, such as a radioactive, chromogenic, or fluorescent signal, that can be used to quantify the amount of bound detectable moiety in a sample. Quantitation of the signal is achieved by, *e.g.*, scintillation counting, densitometry, or flow cytometry.

[0074] The term "activity" as used in the application refers to the biological functions of a molecule, such as a protein encoded by a gene of interest, *e.g.*, osteopontin or EpCAM. This term encompasses biological functions such as enzymatic activity, specific interaction with other molecules, regulatory effects on biological events at molecular or cellular level, and the like.

[0075] The term "inhibiting" or "inhibition" as used herein refers to a negative regulatory effect on the function or activity of an intended target molecule, such that the function or activity, *e.g.*, enzymatic activity or specific interaction with other molecules, is detectably diminished or effectively abolished.

[0076] The term "antagonist" as used herein refers to a compound that is capable of negatively regulating the biological activity of a target molecule, *e.g.*, osteopontin or EpCAM. An antagonist may effectuate the negative regulation by various means, such as by suppression of the expression of the target gene at transcriptional or translational level, or by interfering with the target molecule in its specific interaction with other molecules.

[0077] The term "antisense" as used in the context of describing a polynucleotide, refers to a single-stranded nucleic acid having a nucleotide sequence complementary to at least a portion of a target nucleic acid that encodes a protein of interest (*e.g.*, osteopontin, or EpCAM), or the "sense" sequence. Complementarity between two single-stranded polynucleotides is based on the "A-T G-C" base-pairing rule. For example, the sequence "5'-AGAT-3'," is complementary to the sequence "5'-ATCT-3'". Complementarity between a target nucleic acid and its antisense polynucleotide is typically 100%, i.e., all bases of the antisense polynucleotide match the with the bases of the target nucleic acid, but may be of varying degrees, i.e., there are may be some mis-matched bases. The degree of complementarity between a target nucleic acid and its antisense polynucleotide has significant effects on the efficiency and strength of hybridization. An "antisense" polynucleotide sequence in the present application may correspond to a coding portion (i.e., exon) or a non-coding portion (i.e., intron) of the target nucleic acid.

BRIEF DESCRIPTION OF THE DRAWINGS

[0078] **Figure 1**. Classification of hepatocellular carcinoma with or without metastasis by gene expression. A) Multidimensional scaling analysis of 50 primary and metastatic HCC samples using 143 significant genes (p<0.0005) from supervised class comparison analysis of all 5 clinical groups, i.e., P, P-M, PT, PT-M, PN. The axes represent the first three principal components of these genes. P, primary HCC with intra-hepatic spreads; P-M, metastatic lesion of P; PT, primary HCC with tumor thrombus in portal vein; PN, metastasis-free primary HCC samples. B) Hierarchical clustering of 30 primary HCC samples from P, PT, and PN groups using 383 significant genes (p<0.0005) derived from supervised class comparison.

[0079] **Figure 2**. Prediction of metastasis and survival with metastasis predictor model derived from "leave-one-out' cross-validated compound covariate predictor classification. A) Metastasis predictor model used in 40 training and testing HCC patients. The predictor was based on a training set (circle) including 10 PN and 10 PT primary HCC samples that were previously used in the compound covariate predictor classification and 20 primary blinded HCC samples that were not used in the training procedure. The predictor uses 153 significant genes that distinguish between these two groups. B) Multidimensional scaling analysis of 40 primary HCC samples using 153 significant genes from the predictor. Patient IDs are indicated. C) Kaplan-Meier survival curves for 40 PN, PT and P patients. Cross marks indicate time of censorship.

[0080] **Figure 3**. Candidate genes associated with metastatic HCC. A) Hierarchical clustering of top 30 candidate genes whose expressions were altered largely in PT and PT-M, but rarely in PN. Each row represents an individual gene and each column represents an individual tumor sample. Genes were ordered by centered correlation and complete linkage according the ratio of its abundance to the median abundance of all genes among all tumor samples. Pseudo colors indicate differential expression: green squares, transcript levels below the median; black squares, transcript levels equal to the median; red squares, transcript levels greater than the median; gray squares, missing data. Dendrogram was based on 10 primary PN (green) and 10 primary PT (red) samples. B) Relative expression ratio of OPN by cDNA microarray analysis in 10 primary PN samples (green bars) and 10 primary PT samples (red bars) with accompanying metastasis (black bars). C and D) Semi-quantitative RT-PCR analysis of OPN mRNA level in primary HCC samples with or without metastasis.

[0081] **Figure 4**. Immunohistochemical analysis of osteopontin in normal liver and hepatocellular carcinoma. Primary tumor cells (tumor S30) show cytoplasmic osteopontin immunoreactivity, especially in the area with high density of vasculature (panels b and d), but fibrous septa region (panels b and d) or normal liver parenchyma cells show no reactivity (panels a and c; normal liver 914). Magnification, x50. (H&E, x50).

[0082] **Figure 5**. Role of osteopontin in promoting HCC metastasis. A) The level of osteopontin of CCL13, SK-Hep-1, and Hep3B cells was determined by Western blotting with a rat monoclonal anti-OPN antibody. A monoclonal ß-actin antibody was used as internal control. Densitometry was used to quantify the amount of OPN, which was normalized to actin. OPN level is indicated as relative folds. B) CCL13, SK-Hep-1 or Hep3B cells were incubated with or without a murine recombinant osteopontin protein or a neutralizing antibody against osteopontin and their invasiveness was determined by the Matrigel Basement Membrane *Cell* Invasion Chamber. Data is an average of triplicate determinants for each condition and is expressed as the mean percent invasion (plus one standard deviation) through the Matrigel Matrix and membrane (matrigel chamber) relative to the migration through the control membrane (control chamber). C) The invasiveness of five additional HCC cell lines (SMMC7721, MHCC97, HuH1, HuH4 and HuH7) through matrigel matrix in responding to osteopontin neutralizing antibody was determined as above.

D) Representative lung tissue sections (H&E stain; magnification x100) from mice at 35 days following s.c. injection of HCCLM3 cells without (upper panel) or with (bottom panel) anti-OPN neutralizing antibody are shown. Arrows indicate the tumor grades. E) Primary tumor size was monitored at various weeks following s.c. injection of HCCLM3 cells into nude mice. Data are an average of 10 mice. F) The formation of pulmonary metastases in nude mice was determined at 35 days following s.c. injection of HCCLM3 cells with or without anti-OPN antibody. The number of metastatic foci was quantified based on their grades. Data are an average of 10 mice per group. The groups with significant p values (<0.05) are indicated by the asterisk.

[0083] **Figure 6**. Potential oncogenic role of EpCAM in HCC development. a) and b) The expression level of EpCAM in various chronic liver disease (CLD) liver samples as analyzed by microarray (a) or RT-PCR (b). c) EpCAM expression in cells from normal human fibroblasts (NHF-hTERT), normal liver (CCL13) and hepatoma (SK-Hep-1, Hep3B, Huh1, Huh4, Huh7, and HepG2) was analyzed by western blotting with a monoclonal antibody against EpCAM. A monoclonal antibody against beta-actin was used as an internal control. d) Cell proliferation of Hep3B, Huh1, and Huh4 cells was determined by MTT assay and data were an average of 3 independent experiments. e) Effective silencing of EpCAM expression by siRNA was determined by western blotting analysis. f) Growth inhibition of Hep3B cells by EpCAM siRNA as determined by MTT assay.

DETAILED DESCRIPTION OF THE INVENTION

[0084] Hepatocellular carcinoma (HCC) is one of the most common and aggressive malignant tumors in the world, with high prevalence especially in Asia and Africa, and relatively low prevalence in Europe and North America (Parkin et al., *CA Cancer J. Clin.* **49**:33-64, 1999; Pisani et al., *Int. J. Cancer* **83**:18-29, 1999). Recent studies indicate that the incidence of HCC in the U.S. and in the U.K. has significantly increased over the last two decades (Taylor-Robinson et al., *Lancet* **350**:1142-1143, 1997; El-Serag and Mason, *N. Eng. J. Med.* **340**:745-750, 1999). Most of the HCC patients are incurable due to their poor prognosis. Although routine screening of individuals who are at the risk for developing HCC may provide an opportunity for some patients with an extended life, many patients are still diagnosed with advanced HCC with little improved survival (*see, e.g.*, Yang et al, *J. Cancer Res. Clin. Oncol.* **123**:357-360, 1997; Izzo et al., *Ann. Surg.* **227**:513-518, 1998). While a small subset of HCC patients qualifies for surgical intervention, the improvement on long-term survival is only modest. The extremely poor prognosis of HCC is largely because of a high rate of recurrence after surgery, or intra-hepatic metastases that develop by invasion of the portal vein or spreading to other parts of the liver, whereas extrahepatic metastases are less common (*see, e.g.*, Genda et al., *Hepatology* **30**:1027-1036, 1999). These data indicate that the liver is the main target organ of HCC metastasis. It has been demonstrated in animal model systems as well as in patients that the portal vein is the main route for intrahepatic metastases of metastatic HCC cells (*see, e.g.*, Mitsunobu et al., *Clin. Exp. Metastasis* **14**:520-529, 1996). This specific feature of HCC underscores the need to develop an accurate molecular profiling model for better diagnosis and therapeutic targets for the treatment of HCC patients with intrahepatic metastases.

[0085] Current studies have largely been focused on individual candidate genes (*see, e.g.*, Osada et al., *Hepatology* **24**:1460-1467, 1996; Guo et al., *Hepatology* **28**:1481-1488, 1998; Hui et al., *Int. J. Cancer* **84**:604-608, 1999), which may be insufficient to reflect the precise biological nature of metastatic HCC. The microarray technology has offered an opportunity to probe disease-related gene expressions at a global genome scale (*see, e.g.*, Schena et al., *Science* **270**:467-470, 1995). This approach has allowed the successful molecular classification of several human malignant tumors in regarding their stage, prognostic outcome, or response to therapy (Alizadeh et al., *Nature* **403**:503-511, 2000; Bittner et al., *Nature* **406**:536-540, 2000; Perou et al., *Nature* **406**:747-752, 2000; Khan et al., *Nat. Med.* **7**:673-679, 2001; Pomeroy et al., *Nature* **415**: 436-442, 2002; Shipp et al., *Nat. Med.* **8**:68-74, 2002). A few reports have dealt with the gene expression profiles of primary HCC samples (Okabe et al., *Cancer Res.* **61**:2129-2137, 2001; Xu et al., *Proc. Natl. Acad. Sci. U.S.A.* **98**:15089-15094, 2001). However, little is known about the molecular signatures associated with a poor prognostic feature of patients with metastatic HCC.

[0086] Using cDNA microarray-based gene expression profiling, the global changes associated with metastasis are investigated. The initial goal was to identify genes that can discriminate primary tumors from their matched intra-hepatic metastatic lesions. It is revealed that intrahepatic metastatic lesions are indistinguishable from their primary tumors, regardless of tumor size, encapsulation, and patient's age, whereas primary metastasis-free HCC is distinct from primary HCC with metastasis. These data indicate that changes favoring intrahepatic metastasis are initiated in the primary HCC. Moreover, an important gene, osteopontin, a secreted phosphoprotein, emerges in HCC metastasis. Osteopontin overexpression correlated with primary HCC with metastatic potential and invasiveness of liver tumor-derived cell lines *in vitro*, and an osteopontin-neutralizing antibody efficiently blocked *in vitro* invasion and *in vivo* pulmonary metastasis of HCC cells. These studies identify osteopontin both as a molecular marker for defining HCC patients with metastatic potential and as a potential therapeutic target for treating metastatic HCC.

[0087] A similar approach is used to develop a gene expression prediction model for the potential to develop HCC in patients with chronic liver diseases. By comparing the gene expression profiles of patients epidemiologically at high risk for developing HCC with the gene expression profile of patients epidemiologically at low risk for developing HCC, cellular markers are identified so as to allow the identification of individuals with chronic liver diseases at high risk for developing HCC. The patients with severe liver diseases include those diagnosed with chronic hepatitis B infection, hepatitis C infection, hemochromatosis, Wilson’s disease, alcoholic liver disease, autoimmune hepatitis, and primary biliary cirrhosis. High risk precancerous diseases include chronic hepatitis B infection, hepatitis C infection, hemochromatosis, and Wilson's disease. Low risk precancerous diseases include alcoholic liver disease, autoimmune hepatitis, and primary biliary cirrhosis. One gene identified to be associated with elevated risk of developing HCC in patients with severe liver diseases is EpCAM. Growth suppression of liver cancer cells has been observed upon inhibition of EpCAM expression, identifying its important role in HCC development and as a therapeutic target for preventing HCC in patients with chronic liver diseases.

[0088] One particular aspect of the invention provides methods for clustering co-regulated genes in patients suspected of having metastatic HCC or the potential to develop HCC into gene expression profiles. This section provides a more detailed discussion of methods for clustering co-regulated genes.

I. DNA MICROARRAY ANALYSIS

A. Gene expression profile Classification by Cluster Analysis

[0089] For many applications of the present invention, it is desirable to find basis gene expression profiles that are co-regulated in the non-metastatic HCC samples, the metastatic HCC samples, the high risk for developing HCC samples and the low risk for developing HCC samples. A preferred embodiment for identifying such basis gene expression profiles involves clustering algorithms (for reviews of clustering algorithms, see, e.g., Fukunaga, 1990, Statistical Pattern Recognition, 2nd Ed., Academic Press, San Diego; Everitt, 1974, Cluster Analysis, London: Heinemann Educ. Books; Hartigan, 1975, Clustering Algorithms, New York: Wiley; Sneath and Sokal, 1973, Numerical Taxonomy, Freeman; Anderberg, 1973, Cluster Analysis for Applications, Academic Press: New York).

[0090] In some embodiments employing cluster analysis, the expression of a large number of genes is monitored in biological samples obtained from different sources A table of data containing the gene expression measurements is used for cluster analysis. Cluster analysis operates on a table of data which has the dimension m x k wherein m is the total number of conditions or perturbations and k is the number of genes measured.

[0091] A number of clustering algorithms are useful for clustering analysis. Clustering algorithms use dissimilarities or distances between objects when forming clusters. In some embodiments, the distance used is Euclidean distance in multidimensional space. The Euclidean distance may be squared to place progressively greater weight on objects that are further apart. Alternatively, the distance measure may be the Manhattan distance. In other embodiments unsupervised hierarchical clustering of a table of data may be performed using the CLUSTER or TREEVIEW software (Eisen et al., *Proc. Natl. Acad. Sci. U.S.A.* **95**:14863-14868, 1998) using median centered correlation and complete linkage.

[0092] Various cluster linkage rules are useful for the methods of the invention. Single linkage, a nearest neighbor method, determines the distance between the two closest objects. By contrast, complete linkage methods determine distance by the greatest distance between any two objects in the different clusters. This method is particularly useful in cases when genes or other cellular constituents form naturally distinct "clumps." Alternatively, the un-weighted pair-group average defines distance as the average distance between all pairs of objects in two different clusters. This method is also very useful for clustering genes or other cellular constituents to form naturally distinct "clumps." Finally, the weighted pair-group average method may also be used. This method is the same as the unweighted pair-group average method except that the size of the respective clusters is used as a weight. This method is particularly useful for embodiments where the cluster size is suspected to be greatly varied (Sneath and Sokal, 1973, Numerical taxonomy, San Francisco. W. H. Freeman & Co.). Other cluster linkage rules, such as the unweighted and weighted pair-group centroid and Ward's method are also useful for some embodiments of the invention. See., e g, Ward, 1963, *J. Am. Stat Assn.* **58**:236; Hartigan, 1975, *Clustering algorithms*, New York: Wiley.

[0093] In one particularly preferred embodiment, the cluster analysis used is the BRB-ArrayTools software, an integrated package for the visualization and statistical analysis of cDNA microarray gene expression data developed by the Biometric Research Branch of the National Cancer Institute, for both unsupervised and supervised analyses. The Class Comparison Tool based on univariate F-tests may be used to find genes differentially expressed between predefined clinical groups at a significance level of P <0.001 or 0.002. The permutation distribution of the F-statistic, based on 2000 random permutations may also used to confirm statistical significance. The multi-variate Compound Covariate Predictor (CCP) Tool with a "leave-one-out" cross-validation test using 2000 random permutations at a significant level of P<0.001 may be used to classify predefined clinical groups based on their gene expression profiles. In each cross-validation step one sample is omitted and a multivariate CCP is created based on the genes that are univariately significant at the specified level in the training set consisting of the samples not omitted. This CCP is used to classify the omitted sample and it is then noted whether the classification is correct or incorrect. This is repeated with all samples excluded one at a time. The total cross-validated misclassification rate is thereby determined. The statistical significance of the cross-validated misclassification rate is determined by repeating the entire cross-validation procedure to data with the class membership labels randomly permuted 2000 times. The CCP is based on a weighted linear combination of gene expression variables that are univariately significant in the training set with the weights being the corresponding t-statistics as described in Radmacher et al., *Journal of Computational Biology*, in press, 2002. An example of a clustering `tree` output is shown in Figures 1 and 3 (see, also, Example 1, infra).

[0094] Gene expression profiles may be defined based on the many smaller branches in the tree, or a small number of larger branches by cutting across the tree at different levels. The choice of cut level may be made to match the number of distinct clinical groups expected. If little or no prior information is available about the number of groups, then the tree should be divided into as many branches as are truly distinct. `Truly distinct` may be defined by a minimum distance value between the individual branches. This distance is the vertical coordinate of the horizontal connector joining two branches (see Figure 1B). Typical values are in the range 0.2 to 0.4 where 0 is perfect correlation and 1 is zero correlation, but may be larger for poorer quality data or fewer experiments in the training set, or smaller in the case of better data and more experiments in the training set.

[0095] Preferably, `truly distinct` may be defined with an objective test of statistical significance for each bifurcation in the tree. In one aspect of the invention, the Compound Covariat Predictor (CCP) tool with "leave one out" cross-validation test using 2000 random permutations at a predefined significant level is used to define an objective test. The distribution of tractional improvements obtained from the CCP procedure is an estimate of the distribution under the null hypothesis that a particular classification is correct or incorrect.

[0096] Another aspect of the cluster analysis method of this invention provides the definition of basis vectors for use in profile projection described in the following sections.

B. Profile Comparison and Classification

[0097] One aspect of the invention provides methods for drug discovery. In one embodiment, gene expression profiles are defined using cluster analysis. The genes within a gene expression profile are indicated as potentially co-regulated under the conditions of interest. Co-regulated genes are further explored as potentially being involved in a regulatory pathway. Identification of genes involved in a regulatory pathway provides useful information for designing and screening new drugs.

[0098] In some embodiments of the invention, drug candidates are screened for their therapeutic activity. In one embodiment, desired drug activity is to affect one particular genetic regulatory pathway. In this embodiment, drug candidates are screened for their ability to affect the gene expression profile corresponding to the regulatory pathway. In another embodiment, a new drug is desired to replace an existing drug. In this embodiment, the projected profiles of drug candidates are compared with that of the existing drug to determine which drug candidate has activities similar to the existing drug.

[0099] In some embodiments, the method of the invention is used to decipher pathway arborization and kinetics. When a receptor is triggered (or blocked) by a ligand, the excitation of the downstream pathways can be different depending on the exact temporal profile and molecular domains of the ligand interaction with the receptor. Simple examples of the differing effects of different ligands are the phenotypical differences that arise between responses to agonists, partial agonists, negative antagonists, and antagonists, and that are expected to occur in response to covalent vs. noncovalent binding and activation of different molecular domains on the receptor. See, Ross, *Pharmacodynamics: Mechanisms of Drug Action and the Relationship between Drug Concentration and Effect in The Pharmacological*

Basis of Therapeutics (Gilman et al. ed., McGraw Hill, New York, 1996) FIG. 4A illustrates two different possible responses of a pathway cascade.

[0100] In some embodiments of the invention, receptors for ligands such as OPN may be investigated using the projection method of the invention to simplify the observed temporal responses to receptor/ligand interactions over the responding genes. In some particularly preferred embodiments, the gene expression profiles and temporal profiles involved are discovered. The profile of temporal responses of a large number of genes are projected onto the predefined gene expression profiles to obtain a projected profile of temporal responses. The projection process simplifies the observed responses so that different temporal responses may be detected and discriminated more accurately.

C. Illustrative Diagnostic Applications

[0101] One aspect of the invention provides methods for diagnosing diseases of humans, animals and plants. Those methods are also useful for monitoring the progression of diseases and the effectiveness of treatments.

[0102] In one embodiment of the invention, a patient cell sample such as a biopsy from a patient's diseased tissue such as metastatic HCC, is assayed for the expression of a large number of genes. The gene expression profile is projected into a profile of gene expression profile expression values according to a definition of gene expression profiles. The projected profile is then compared with a reference database containing reference projected profiles. If the projected profile of the patient matches best with a cancer profile in the database, the patient's diseased tissue is diagnosed as being cancerous. Similarly, when the best match is to a profile of another disease or disorder, a diagnosis of such other disease or disorder is made.

[0103] In another embodiment, a tissue sample is obtained from a patient's tumor. The tissue sample is assayed for the expression of a large number of genes of interest. The gene expression profile is projected into a profile of gene expression profile expression values according to a definition of gene expression profiles. The projected profile is compared with projected profiles previously obtained from the same tumor to identify the change of expression in gene expression profiles. A reference library is used to determine whether the gene expression profile changes indicate tumor progression such as metastasis. A similar method is used to stage other diseases and disorders. Changes of gene expression profile expression values in a profile obtained from a patient under treatment can be used to monitor the effectiveness of the treatment, for example, by comparing the projected profile prior to treatment with that after treatment.

D. Analytic Kit Implementation

[0104] In a preferred embodiment, the methods of this invention can be implemented by use of kits for determining the responses or state of a biological sample. Such kits contain microarrays, such as those described in subsections below. The microarrays contained in such kits comprise a solid phase, *e.g.*, a surface, to which probes are hybridized or bound at a known location of the solid phase. Preferably, these probes consist of nucleic acids of known, different sequence, with each nucleic acid being capable of hybridizing to an RNA species or to a cDNA species derived therefrom. In particular, the probes contained in the kits of this invention are nucleic acids capable of hybridizing specifically to nucleic acid sequences derived from RNA species which are known to increase or decrease in response to perturbations to the particular protein whose activity is determined by the kit. The probes contained in the kits of this invention preferably substantially exclude nucleic acids which hybridize to RNA species that are not increased in response to perturbations to the particular protein whose activity is determined by the kit, such as osteopontin.

[0105] In a preferred embodiment, a kit of the invention also contains a database of gene expression profile definitions such as the databases described above or an access authorization to use the database described above from a remote networked computer.

[0106] In another preferred embodiment, a kit of the invention further contains expression profile projection and analysis software capable of being loaded into the memory of a computer system such as the one described supra in the subsection, and illustrated in Example 1. The expression profile analysis software contained in the kit of this invention, is essentially identical to the expression profile analysis software described above in Example 1.

[0107] Alternative kits for implementing the analytic methods of this invention will be apparent to one of skill in the art and are intended to be comprehended within the accompanying claims. In particular, the accompanying claims are intended to include the alternative program structures for implementing the methods of this invention that will be readily apparent to one of skill in the art.

E. Methods for Determining Biological Response Profiles

[0108] This invention utilizes the ability to measure the responses of a biological system to a large variety of perturbations. This section provides some exemplary methods for measuring biological responses. One of skill in the art would appreciate that this invention is not limited to the following specific methods for measuring the responses of a biological system.

1. Transcript Assay Using DNA Array

[0109] This invention is particularly useful for the analysis of gene expression profiles. One aspect of the invention provides methods for defining co-regulated gene expression profiles based upon the correlation of gene expression. Some embodiments of this invention are based on measuring the transcriptional rate of genes.

[0110] The transcriptional rate can be measured by techniques of hybridization to arrays of nucleic acid or nucleic acid mimic probes, described in the next section, or by other gene expression technologies, such as those described in the subsequent subsection. However measured, the result is either the absolute, relative amounts of transcripts or response data including values representing RNA abundance ratios, which usually reflect DNA expression ratios (in the absence of differences in RNA degradation rates).

[0111] In various alternative embodiments of the present invention, aspects of the biological state other than the transcriptional state, such as the translational state, the activity state, or mixed aspects can be measured.

[0112] Preferably, measurement of the transcriptional state is made by hybridization to DNA microarrays, which are described in this section. Certain other methods of transcriptional state measurement are described later in this subsection.

[0113] In a preferred embodiment the present invention makes use of DNA microarrays. DNA microarrays can be employed for analyzing the transcriptional state in a biological sample and especially for measuring the transcriptional states of a biological sample exposed to graded levels of a drug of interest or to graded perturbations to a biological pathway of interest.

[0114] In one embodiment, DNA microarrays are produced by hybridizing detectably labeled polynucleotides representing the mRNA transcripts present in a cell (e.g., fluorescently labeled cDNA synthesized from total cell mRNA) to a microarray. A microarray is a surface with an ordered array of binding (e.g., hybridization) sites for products of many of the genes in the genome of a cell or organism, preferably most or almost all of the genes. Microarrays can be made in a number of ways, of which several are described below. However produced microarrays share certain preferred characteristics: The arrays are reproducible, allowing multiple copies of a given array to be produced and easily compared with each other. Preferably the microarrays are small, usually smaller than 5^2 cm, and they are made from materials that are stable under binding (e.g., nucleic acid hybridization) conditions. A given binding site or unique set of binding sites in the microarray will specifically bind the product of a single gene in the cell. Although there may be more than one physical binding site (hereinafter "site") per specific mRNA, for the sake of clarity the discussion below will assume that there is a single site.

[0115] It will be appreciated that when cDNA complementary to the RNA of a cell is made and hybridized to a microarray under suitable hybridization conditions, the level of hybridization to the site in the array corresponding to any particular gene will reflect the prevalence in the cell of mRNA transcribed from that gene. For example, when detectably labeled (e.g., with a fluorophore) cDNA complementary to the total cellular mRNA is hybridized to a microarray, the site on the array corresponding to a gene (i.e., capable of specifically binding the product of the gene) that is not transcribed in the cell will have little or no signal (e.g., fluorescent signal), and a gene for which the encoded mRNA is prevalent will have a relatively strong signal.

[0116] In preferred embodiments, cDNAs from two different cells are hybridized to the binding sites of the microarray. In the case of drug responses one biological sample is exposed to a drug and another biological sample of the same type is not exposed to the drug. In the case of pathway responses one cell is exposed to a pathway perturbation and another cell of the same type is not exposed to the pathway perturbation. The cDNA derived from each of the two cell types are differently labeled so that they can be distinguished. In one embodiment, for example, cDNA from a cell treated with a drug (or exposed to a pathway perturbation) is synthesized using a fluorescein-labeled dNTP, and cDNA from a second cell, not drug-exposed, is synthesized using a rhodamine-labeled dNTP. When the two cDNAs are mixed and hybridized to the microarray, the relative intensity of signal from each cDNA set is determined for each site on the array, and any relative difference in abundance of a particular mRNA detected.

[0117] In the example described above, the cDNA from the drug-treated (or pathway perturbed) cell will fluoresce green when the fluorophore is stimulated and the cDNA from the untreated cell will fluoresce red. As a result, when the drug treatment has no effect, either directly or indirectly, on the relative abundance of a particular mRNA in a cell, the mRNA will be equally prevalent in both cells and, upon reverse transcription, red-labeled and green-labeled cDNA will be equally prevalent. When hybridized to the microarray, the binding site(s) for that species of RNA will emit wavelengths characteristic of both fluorophores (and appear brown in combination). In contrast, when the drug-exposed cell is treated with a drug that, directly or indirectly, increases the prevalence of the mRNA in the cell, the ratio of green to red fluorescence will increase. When the drug decrease the mRNA prevalence, the ratio will decrease.

[0118] The use of a two-color fluorescence labeling and detection scheme to define alterations in gene expression has been described in, *e.g.*, Shena et al., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* **270**:467-470, 1995, which is incorporated by reference in its entirety for all purposes. An advantage of using cDNA labeled with two different fluorophores is that a direct and internally controlled comparison of the mRNA levels corresponding to each arrayed gene in two cell states can be made, and variations due to minor differences in experimental conditions (e.g., hybridization conditions) will not affect subsequent analyses. However, it will be recognized that it is also possible to use cDNA from a single cell, and compare, for example, the absolute amount of a particular mRNA in, e.g., a drug-treated or pathway-perturbed cell and an untreated cell.

2. Preparation of Microarrays

[0119] Microarrays are known in the art and consist of a surface to which probes that correspond in sequence to gene products (e.g., cDNAs, mRNAs, cRNAs, polypeptides, and fragments thereof), can be specifically hybridized or bound at a known position. In one embodiment, the microarray is an array (i.e., a matrix) in which each position represents a discrete binding site for a product encoded by a gene (e.g., a protein or RNA), and in which binding sites are present for products of most or almost all of the genes in the organism's genome. In a preferred embodiment, the "binding site" (hereinafter, "site") is a nucleic acid or nucleic acid analogue to which a particular cognate cDNA can specifically hybridize. The nucleic acid or analogue of the binding site can be, e.g., a synthetic oligomer, a full-length cDNA, a less-than full length cDNA, or a gene fragment.

[0120] Although in a preferred embodiment the microarray contains binding sites for products of all or almost all genes in the target organism's genome, such comprehensiveness is not necessarily required. Usually the microarray will have binding sites corresponding to at least about 50% of the genes in the genome, often at least about 75%, more often at least about 85%, even more often more than about 90%, and most often at least about 99%. Preferably, the microarray has binding sites for genes relevant to the action of a drug of interest or in a biological pathway of interest. A "gene" is identified as an open reading frame (ORF) of preferably at least 50, 75, or 99 amino acids from which a messenger RNA is transcribed in the organism (e.g., if a single cell) or in some cell in a multicellular organism. The number of genes in a genome can be estimated from the number of mRNAs expressed by the organism, or by extrapolation from, a well-characterized portion of the genome. When the genome of the organism of interest has been sequenced, the number of ORFs can be determined and mRNA coding regions identified by analysis of the DNA sequence. For example, the Saccharomyces cerevisiae genome has been completely sequenced and is reported to have approximately 6275 open reading frames (ORFs) longer than 99 amino acids. Analysis of these ORFs indicates that there are 5885 ORFs that are likely to specify protein products (Goffeau et al., 1996, Life with 6000 genes, Science 274:546-567, which is incorporated by reference in its entirety for all purposes). In contrast, the human genome is estimated to contain approximately $5x10^4$ genes.

3. Preparing Nucleic Acids for Microarrays

[0121] As noted above, the "binding site" to which a particular cognate cDNA specifically hybridizes is usually a nucleic acid or nucleic acid analogue attached at that binding site. In one embodiment, the binding sites of the microarray are DNA polynucleotides corresponding to at least a portion of each gene in an organism's genome. These DNAs can be obtained by, e.g., polymerase chain reaction (PCR) amplification of gene segments from genomic DNA, cDNA (e.g., by RT-PCR), or cloned sequences. PCR primers are chosen, based on the known sequence of the genes or cDNA, that result in amplification of unique fragments (i.e., fragments that do not share more than 10 bases of contiguous identical sequence with any other fragment on the microarray). Computer programs are useful in the design of primers with the required specificity and optimal amplification properties. See, e.g., Oligo version

5.0 (National Biosciences). In the case of binding sites corresponding to very long genes, it will sometimes be desirable to amplify segments near the 3' end of the gene so that when oligo-dT primed cDNA probes are hybridized to the microarray, less-than-full length probes will bind efficiently. Typically each gene fragment on the microarray will be between about 50 bp and about 2000 bp, more typically between about 100 bp and about 1000 bp, and usually between about 300 bp and about 800 bp in length. PCR methods are well known and are described, for example, in Innis et al. eds., 1990, PCR Protocols: A Guide to Methods and Applications, Academic Press Inc., San Diego, Calif., which is incorporated by reference in its entirety for all purposes. It will be apparent that computer controlled robotic systems are useful for isolating and amplifying nucleic acids.

[0122] An alternative means for generating the nucleic acid for the microarray is by synthesis of synthetic polynucleotides or oligonucleotides, e.g., using N-phosphonate or phosphoramidite chemistries (Froehler et al., 1986, Nucleic Acid, Res 14:5399-5407; McBride et al., 1983, Tetrahedron Lett. 24:245-248). Synthetic sequences are between about 15 and about 500 bases in length, more typically between about 20 and about 50 bases. In some embodiments, synthetic nucleic acids include non-natural bases, e.g., inosine. As noted above, nucleic acid analogues may be used as binding sites for hybridization. An example of a suitable nucleic acid analogue is peptide nucleic acid (see, e.g., Egholm et al., 1993, PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules, Nature 365:566-568; see also U.S. Pat. No. 5,539,083).

[0123] In an alternative embodiment, the binding (hybridization) sites are made from plasmid or phage clones of genes, cDNAs (e.g., expressed sequence tags), or inserts therefrom (Nguyen et al., 1995, Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones, Genomics 29:207-209). In yet another embodiment, the polynucleotide of the binding sites is RNA.

4. Attaching Nucleic Acids to the Solid Surface

[0124] The nucleic acid or analogue are attached to a solid support, which may be made from glass, plastic (e.g., polypropylene, nylon), polyacrylamide, nitrocellulose, or other materials. A preferred method for attaching the nucleic acids to a surface is by printing on glass plates, as is described generally by Schena et al., 1995, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270:467-470. This method is especially useful for preparing microarrays of cDNA. See also DeRisi et al., 1996, Use of a cDNA microarray to analyze gene expression patterns in human cancer, Nature Genetics 14:457-460; Shalon et al., 1996, A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, Genome Res. 6:639-645; and Schena et al., 1995, Parallel human genome analysis; microarray-based expression of 1000 genes, Proc. Natl. Acad. Sci. USA 93:10539-11286.

[0125] A second preferred method for making microarrays is by making high-density oligonucleotide arrays. Techniques are known for producing arrays containing thousands of oligonucleotides complementary to defined sequences, at defined locations on a surface using photolithographic techniques for synthesis in situ (see, Fodor et al., 1991, Light-directed spatially addressable parallel chemical synthesis, Science 251:767-773; Pease et al., 1994, Light-directed oligonucleotide arrays for rapid DNA sequence analysis, Proc. Natl. Acad. Sci. USA 91:5022-5026; Lockhart et al., 1996, Expression monitoring by hybridization to high-density oligonucleotide arrays, Nature Biotech 14:1675; U.S. Pat. Nos. 5,578,832; 5,556,752; and 5.510,270, each of which is incorporated by reference in its entirety for all purposes) or other methods for rapid synthesis and deposition of defined oligonucleotides (Blanchard et al., 1996, High-Density, Oligonucleotide arrays, Biosensors & Bioelectronics 11: 687-90). When these methods are used, oligonucleotides (e.g., 20-mers) of known sequence are synthesized directly on a surface such as a derivatized glass slide. Usually, the array produced contains multiple probes against each target transcript. Oligonucleotide probes can be chosen to detect alternatively spliced mRNAs or to serve as various type of control.

[0126] Another preferred method of making microarrays is by use of an inkjet printing process to synthesize oligonucleotides directly on a solid phase.

[0127] Other methods for making microarrays, *e.g.*, by masking (Maskos and Southern, 1992, Nuc. Acids Res. 20:1679-1684), may also be used. In principal, any type of array, for example, dot blots on a nylon hybridization membrane (see Sambrook and Russell, *Molecular Cloning: A Laboratory Manual* 3d ed, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 2001), could be used, although, as will be recognized by those of skill in the art, very small arrays will be preferred because hybridization volumes will be smaller.

5. Generating Labeled Probes

[0128] Methods for preparing total and poly(A)+ RNA are well known and are described generally in Sambrook et al., supra. In one embodiment, RNA is extracted from biological samples of the various types of interest in this invention using guanidinium thiocyanate lysis followed by CsCl centrifugation (Chirgwin et al., 1979, Biochemistry 18:5294-5299). Alternatively, total RNA may be extracted from samples using TRIzol reagent (Life Technologies) according to manufacturer's directions. Poly(A)+ RNA is selected by selection with oligo-dT cellulose (see Sambrook and Russell, *supra*). Biological samples of interest include normal liver samples, non-cancerous liver samples and samples from defined clinical specimens.

[0129] Labeled cDNA is prepared from mRNA by oligo dT-primed or random-primed reverse transcription, both of which are well known in the art (see, e.g., Klug and Berger, 1987, Methods Enzymol. 152:316-325). Reverse transcription may be carried out in the presence of a dNTP conjugated to a detectable label, most preferably a fluorescently labeled dNTP. Alternatively, isolated mRNA can be converted to labeled antisense RNA synthesized by in vitro transcription of double-stranded cDNA in the presence of labeled dNTPs (Lockhart et al., 1996, Expression monitoring by hybridization to high-density oligonucleotide arrays, Nature Biotech. 14:1675, which is incorporated by reference in its entirety for all purposes). In alternative embodiments, the cDNA or RNA probe can be synthesized in the absence of detectable label and may be labeled subsequently, e.g., by incorporating biotinylated dNTPs or rNTP, or some similar means (e.g., photo-cross-linking a psoralen derivative of biotin to RNAs), followed by addition of labeled streptavidin (e.g., phycoerythrin-conjugated streptavidin) or the equivalent.

[0130] When fluorescently-labeled probes are used, many suitable fluorophores are known, including fluorescein, lissamine, phycoerythrin, rhodamine (Perkin Elmer Cetus), Cy2, Cy3, Cy3.5, Cy5, Cy5.5, Cy7, FluorX (Amersham) and others (see, e.g., Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press San Diego, Calif.). It will be appreciated that pairs of fluorophores are chosen that have distinct emission spectra so that they can be easily distinguished.

[0131] In another embodiment, a label other than a fluorescent label is used. For example, a radioactive label, or a pair of radioactive labels with distinct emission spectra, can be used (see Zhao et al., 1995, High density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression, Gene 156:207; Pietu et al., 1996, Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array, Genome Res. 6:492). However, because of scattering of radioactive particles, and the consequent requirement for widely spaced binding sites, use of radioisotopes is a less-preferred embodiment.

[0132] In one embodiment, labeled cDNA is synthesized by incubating a mixture containing 0.5 mM dGTP, dATP and dCTP plus 0.1 mM dTTP plus fluorescent deoxyribonucleotides (e.g., 0.1 mM Rhodamine 110 UTP (Perken Elmer Cetus) or 0.1 mM Cy3 dUTP (Amersham)) with reverse transcriptase (e.g., SuperScript.TM.II, LTI Inc.) at 42°C for 60 minutes.

6. Hybridization to Microarrays

[0133] Nucleic acid hybridization and wash conditions are optimally chosen so that the probe "specifically binds" or "specifically hybridizes" to a specific array site, i.e., the probe hybridizes, duplexes or binds to a sequence array site with a complementary nucleic acid sequence but does not hybridize to a site with a non-complementary nucleic acid sequence. As used herein, one polynucleotide sequence is considered complementary to another when, if the shorter of the polynucleotides is less than or equal to 25 bases, there are no mismatches using standard base-pairing rules or, if the shorter of the polynucleotides is longer than 25 bases, there is no more than a 5% mismatch. Preferably, the polynucleotides are perfectly complementary (no mismatches). It can easily be demonstrated that specific hybridization conditions result in specific hybridization by carrying out a hybridization assay including negative controls (*see, e.g.*, Shalon et al., *supra*, and Chee et al., *supra*).

[0134] Optimal hybridization conditions will depend on the length (*e.g.*, oligomer versus polynucleotide greater than 200 bases) and type (*e.g.*, RNA, DNA, PNA) of labeled probe and immobilized polynucleotide or oligonucleotide. General parameters for specific (i.e., stringent) hybridization conditions for nucleic acids are described in Sambrook et al, supra, and in Ausubel et al., 1987, Current Protocols in Molecular Biology, Greene Publishing and Wiley-Interscience, New York. When the cDNA microarrays of Schena et al. are used, typical hybridization conditions are hybridization in 5xSSC plus 0.2% SDS at 65°C. for 4 hours followed by washes at 25°C. in low stringency wash buffer (1xSSC plus 0.2% SDS) followed by 10 minutes at 25°C. in high stringency wash buffer (0.1xSSC plus 0.2% SDS) (Shena et al., 1996, Proc. Natl. Acad. Sci. USA, 93:10614). Useful hybridization conditions are also provided in, e.g., Tijessen, 1993, Hybridization With Nucleic Acid Probes, Elsevier Science Publishers B. V. and Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press San Diego, Calif.

7. Signal Detection and Data Analysis

[0135] When fluorescently labeled probes are used, the fluorescence emissions at each site of a transcript array can be detected by scanning confocal laser microscopy. Preferably the fluorescent intensities are measured by the Axon GenePix 4000 scanner. In one embodiment, a separate scan, using the appropriate excitation line, is carried out for each of the two fluorophores used. Alternatively, a laser can be used that allows simultaneous specimen illumination at wavelengths specific to the two fluorophores and emissions from the two fluorophores can be analyzed simultaneously (see Shalon et al., 1996, A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, Genome Research 6:639-645, which is incorporated by reference in its entirety for all purposes). In a preferred embodiment, the arrays are scanned with a laser fluorescent scanner with a computer controlled X-Y stage and a microscope objective. Sequential excitation of the two fluorophores is achieved with a multi-line, mixed gas laser and the emitted light is split by wavelength and detected with two photomultiplier tubes. Fluorescence laser scanning devices are described in Schena et al., 1996, Genome Res. 6:639-645 and in other references cited herein. Alternatively, the fiber-optic bundle described by Ferguson et al., 1996, Nature Biotech. 14:1681-1684, may be used to monitor mRNA abundance levels at a large number of sites simultaneously.

[0136] Signals are recorded and, in a preferred embodiment, analyzed by computer, e.g., using a 12 bit analog to digital board. In one embodiment the scanned image is despeckled using a graphics program (*e.g.*, Hijaak Graphics Suite) and then analyzed using an image gridding program that creates a spreadsheet of the average hybridization at each wavelength at each site. If necessary, an experimentally determined correction for "cross talk" (or overlap) between the channels for the two fluors may be made. In a preferred embodiment, the fluorescent intensities were analyzed by the GenePix Pro 3.0 software to subtract the background signals. The expression data were then filtered based on their channel intensities, spots size and flag (missing data) , and the Cy5/Cy3 ratios were calculated and normalized by median-centering the log-ratio of all genes in each array. For any particular hybridization site on the transcript array, a ratio of the emission of the two fluorophores can be calculated. The ratio is independent of the absolute expression level of the cognate gene, but is useful for genes whose expression is significantly modulated by drug administration, gene deletion, or any other tested event.

[0137] According to the method of the invention, the relative abundance of an mRNA in two biological samples is scored as a perturbation and its magnitude determined (i.e., the abundance is different in the two sources of mRNA tested), or as not perturbed (i.e., the relative abundance is the same). In various embodiments, a difference between the two sources of RNA of at least a factor of about 25% (RNA from one source is 25% more abundant in one source than the other source), more usually about 50%, even more often by a factor of about 2 (twice as abundant), 3 (three times as abundant) or 5 (five times as abundant) is scored as a perturbation.

[0138] Preferably, in addition to identifying a perturbation as positive or negative, it is advantageous to determine the magnitude of the perturbation. This can be carried out, as noted above, by calculating the ratio of the emission of the two fluorophores used for differential labeling, or by analogous methods that will be readily apparent to those of skill in the art.

8. Pathway Response and Gene expression profiles

[0139] In one embodiment of the present invention, gene expression profiles are determined by observing the gene expression profile of clinical sample of interest. In one embodiment of the invention, DNA microarrays reflecting the transcriptional state of a biological sample of interest are made by hybridizing a mixture of two differently labeled probes each corresponding (i.e., complementary) to the mRNA of a clinical sample of interest or a reference sample, to the microarray. According to the present invention, the two samples are of the same type, i.e., of the same species and tissue type, but may differ in clinical diagnosis. The genes whose expression are highly correlated may belong to a gene expression profile.

[0140] Further, it is preferable in order to reduce experimental error to reverse the fluorescent labels in two-color differential hybridization experiments to reduce biases peculiar to individual genes or array spot locations. In other words, it is preferable to first measure gene expression with one labeling (*e.g.*, labeling perturbed cells with a first fluorochrome and unperturbed cells with a second fluorochrome) of the mRNA from the two cells being measured, and then to measure gene expression from the two cells with reversed labeling (*e.g.*, labeling perturbed cells with the second fluorochrome and unperturbed cells with the first fluorochrome). Multiple measurements over exposure levels and perturbation control parameter levels provide additional experimental error control. With adequate sampling a trade-off may be made when choosing the width of the spline function S used to interpolate response data between averaging of errors and loss of structure in the response functions.

9. Other Methods of Transcriptional State Measurement

[0141] The transcriptional state of a cell may be measured by other gene expression technologies known in the art. Several such technologies produce pools of restriction fragments of limited complexity for electrophoretic analysis, such as methods combining double restriction enzyme digestion with phasing primers (see, e.g., European Patent O 534858 A1, filed Sep. 24, 1992, by Zabeau et al.), or methods selecting restriction fragments with sites closest to a defined mRNA end (see, e.g., Prashar et al., 1996, Proc. Natl. Acad. Sci. USA 93:659-663). Other methods statistically sample cDNA pools, such as by sequencing sufficient bases (e.g., 20-50 bases) in each of multiple cDNAs to identify each cDNA, or by sequencing short tags (e.g., 9-10 bases) which are generated at known positions relative to a defined mRNA end (see, e.g, Velculescu, 1995, Science 270:484-487).

10. Measurement of Other Aspects of Biological State

[0142] In various embodiments of the present invention, aspects of the biological state other than the transcriptional state, such as the translational state, the activity state, or mixed aspects can be measured in order to obtain drug and pathway responses. Details of these embodiments are described infra.

11. Embodiments Based on Translational State Measurements.

[0143] Measurement of the translational state may be performed according to several methods. For example, whole genome monitoring of protein (i.e., the "proteome," Goffeau et al., supra) can be carried out by constructing a microarray in which binding sites comprise immobilized, preferably monoclonal, antibodies specific to a plurality of protein species encoded by the cell genome. Preferably, antibodies are present for a substantial fraction of the encoded proteins, or at least for those proteins relevant to the action of a drug of interest. Methods for making monoclonal antibodies are well known (see, e.g., Harlow and Lane, 1988, *Antibodies: A Laboratory Manual*, Cold Spring Harbor, N.Y. which is incorporated in its entirety for all purposes). In a preferred embodiment, monoclonal antibodies are raised against synthetic peptide fragments designed based on genomic sequence of the cell. With such an antibody array, proteins from the cell are contacted to the array and their binding is assayed with assays known in the art.

[0144] Alternatively, proteins can be separated by two-dimensional gel electrophoresis systems. Two-dimensional gel electrophoresis is well-known in the art and typically involves iso-electric focusing along a first dimension followed by SDS-PAGE electrophoresis along a second dimension. See, e.g., Hames et at., 1990, Gel Electrophoresis of Proteins: A Practical Approach, IRL Press, New York; Shevchenko et al., 1996, Proc. Nat'l Acad. Sci. USA 93:1440-1445; Sagliocco et al., 1996, Yeast 12:1519-1533; Lander, 1996, Science 274:536-539. The resulting electropherograms can be analyzed by numerous techniques, including mass spectrometric techniques, western blotting and immunoblot analysis using polyclonal and monoclonal antibodies, and internal and N-terminal micro-sequencing. Using these techniques, it is possible to identify a substantial fraction of all the proteins produced under given physiological conditions, including in cells (*e.g.*, in yeast) exposed to a drug, or in cells modified by, *e.g.*, deletion or over-expression of a specific gene.

12. Embodiments Based on Other Aspects of the Biological State

[0145] Even though methods of this invention are illustrated by embodiments involving gene expression profiles, the methods of the invention are applicable to any cellular constituent that can be monitored.

[0146] In particular, where activities of proteins relevant to the characterization of a perturbation, such as drug action, can be measured, embodiments of this invention can be based on such measurements. Activity measurements can be performed by any functional, biochemical, or physical means appropriate to the particular activity being characterized. Where the activity involves a chemical transformation, the cellular protein can be contacted with the natural substrate(s), and the rate of transformation measured. Where the activity involves association in multimeric units, for example association of an activated DNA binding complex with DNA, the amount of associated protein or secondary consequences of the association, such as amounts of mRNA transcribed, can be measured. Also, where only a functional activity is known, for example, as in cell cycle control, performance of the function can be observed. However known and measured, the changes in protein activities form the response data analyzed by the foregoing methods of this invention.

[0147] In alternative and non-limiting embodiments, response data may be formed of mixed aspects of the biological state of a cell. Response data can be constructed from. e.g., changes in certain mRNA abundances, changes in certain protein abundances, and changes in certain protein activities.

II. Proteomic Analysis

[0148] In another aspect, the invention provides methods for detecting markers which are differentially present in the samples of a metastatic HCC tumor or tissue samples of patients predisposed for HCC (*e.g.*, patients at high risk for developing HCC but where the tumor is undetectable). The markers can be detected in a number of biological samples. The sample is preferably a biological tissue sample lysate.

[0149] Any suitable methods can be used to detect one or more of the markers described herein. For example, gas phase ion spectrometry can be used. This technique includes, *e.g.*, laser desorption/ionization mass spectrometry. Preferably, the sample is prepared prior to gas phase ion spectrometry, *e.g.*, pre-fractionation, two-dimensional gel chromatography, high performance liquid chromatography, *etc.* to assist detection of markers. Detection of markers can be achieved using methods other than gas phase ion spectrometry. For example, immunoassays can be used to detect the markers in a sample. These detection methods are described in detail below.

A. Detection by Gas Phase Ion Spectrometry

[0150] Markers present in a biological sample can be detected using gas phase ion spectrometry, and preferably, mass spectrometry. In one embodiment, matrix-assisted laser desorption/ionization ("MALDI") mass spectrometry can be used. In another embodiment, surface-enhanced laser desorption/ionization mass spectrometry ("SELDI") can be used.

1. Preparation of a Sample Prior to Gas Phase Ion Spectrometry

[0151] One or combination of standard techniques well known in the art can be used to prepare a sample to further assist detection and characterization of markers in a sample. For example, a sample can be pre-fractionated to provide a less complex biological sample prior to gas phase ion spectrometry analysis using one or more of the following methods: size exclusion chromatography, Anion Exchange Chromatography, Affinity Chromatography, Sequential Extraction, Gel Electrophoresis, high performance liquid chromatography (HPLC).

[0152] Optionally, a marker can be modified before analysis to improve its resolution or to determine its identity. For example, the markers may be subject to proteolytic digestion before analysis. Fragments from a digestion by a suitable protease, such as trypsin, may function as a fingerprint for the markers, thereby enabling their detection indirectly.

2. Contacting a Sample with a Substrate for Gas Phase Ion Spectrometry Analysis

[0153] A biological sample can be contacted with a substrate, such as a spectrometer probe adapted for use with a gas phase ion spectrometer. Alternatively, a substrate can be a separate material that can be placed onto a spectrometer probe that is adapted for use with a gas phase ion spectrometer.

[0154] A spectrometer probe can be in any suitable shape as long as it is adapted for use with a gas phase ion spectrometer (*e.g.*, removably insertable into a gas phase ion spectrometer). The spectrometer probe substrate can be made of any suitable material, solid or porous. Spectrometer probes suitable for use in embodiments of the invention are described in, *e.g.*, U.S. Patent No. 5,617,060 (Hutchens and Yip) and WO 98/59360 (Hutchens and Yip).

[0155] If complexity of a sample has been substantially reduced as described above, the sample can be contacted with any suitable substrate for gas phase ion spectrometry. Prior to gas phase ions spectrometry analysis, an energy absorbing molecule ("EAM") or a matrix material is typically applied to markers on the substrate surface. The energy absorbing molecule and the sample containing markers can be contacted in any suitable manner.

[0156] Complexity of a sample can be further reduced using a substrate that comprises adsorbents capable of binding one or more markers. Adsorbents that bind the markers can be applied to the substrate in any suitable pattern (*e.g.*, continuous or discontinuous), and a sample can be contacted with a substrate comprising an adsorbent in any suitable manner, *e.g.*, bathing, soaking, dipping, spraying, washing over, or pipetting, *etc.* Following the contact, it is preferred that unbound materials on the substrate surface are washed out so that only the bound materials remain on the substrate surface.

3. Desorption/Ionization and Detection

[0157] Markers on the substrate surface can be desorbed and ionized using gas phase ion spectrometry. Any suitable gas phase ion spectrometers can be used as long as it allows markers on the substrate to be resolved. Preferably, gas phase ion spectrometers allow quantitation of markers. In one embodiment, the gas phase ion spectrometer is a mass spectrometer, preferably a laser desorption time-of-flight mass spectrometer. In another embodiment, an ion mobility spectrometer can be used to detect markers. In yet another embodiment, a total ion current measuring device can be used to detect and characterize markers.

4. Analysis of Data

[0158] Data generated by desorption and detection of markers can be analyzed using any suitable means. In one embodiment, data sets are analyzed with the use of a programmable digital computer. The computer program generally contains a readable medium that stores codes. Certain code can be devoted to memory that includes the location of each feature on a spectrometer probe, the identity of the adsorbent at that feature and the elution conditions used to wash the adsorbent. The computer also contains code that receives as input, data on the strength of the signal at various molecular masses received from a particular addressable location on the spectrometer probe. These data can indicate the number of markers detected, including the strength of the signal generated by each marker.

[0159] Data analysis can include the steps of determining signal strength (*e.g.*, height of peaks) of a marker detected and removing "outerliers" (data deviating from a predetermined statistical distribution). The observed peaks can be normalized, a process whereby the height of each peak relative to some reference is calculated. For example, a reference can be background noise generated by instrument and chemicals (*e.g.*, energy absorbing molecule) which is set as zero in the scale. Then the signal strength detected for each marker or other biomolecules can be displayed in the form of relative intensities in the scale desired (*e.g.*, 100). Alternatively, a standard (*e.g.*, a serum protein) may be admitted with the sample so that a peak from the standard can be used as a reference to calculate relative intensities of the signals observed for each marker or other markers detected.

[0160] The computer can transform the resulting data into various formats for displaying. In one format, referred to as "spectrum view or retentate map," a standard spectral view can be displayed, wherein the view depicts the quantity of marker reaching the detector at each particular molecular weight. In another format, referred to as "peak map," only the peak height and mass information are retained from the spectrum view, yielding a cleaner image and enabling markers with nearly identical molecular weights to be more easily seen. In yet another format, referred to as "gel view," each mass from the peak view can be converted into a grayscale image based on the height of each peak, resulting in an appearance similar to bands on electrophoretic gels. In yet another format, referred to as "3-D overlays," several spectra can be overlaid to study subtle changes in relative peak heights. In yet another format, referred to as "difference map view," two or more spectra can be compared, conveniently highlighting unique markers and markers which are up- or down-regulated between samples. Marker profiles (spectra) from any two samples may be compared visually. In yet another format, Spotfire Scatter Plot can be used, wherein markers that are detected are plotted as a dot in a plot, wherein one axis of the plot represents the apparent molecular of the markers detected and another axis represents the signal intensity of markers detected. For each biological sample, markers that are detected and the amount of markers present in the biological sample can be saved in a computer readable medium. These data can then be compared to a control (*e.g.*, a profile or quantity of markers detected in control, *e.g.*, patients in whom metastatic HCC or tissue samples of someone predisposed for HCC is undetectable).

[0161] A method for predicting the potential of developing metastasis in an HCC patient or developing HCC in a patient with chronic liver disease can be embodied by code that is executed by a digital computer capable of processing data sets derived from signals from arrays after contact with patient samples. The code can be executed by the digital computer to created an analytical model. The code may be stored on any suitable computer readable media. Examples of computer readable media include magnetic, electronic, or optical disks, tapes, sticks, chips, etc. The code may also be written in any suitable computer programming language including, visual basis, Fortran, C, C^{++}, etc. The digital computer may be a micro, mini, or large frame computer using any standard or specialized operating system such as a Windows™ based operating system. A standard PC (personal computer) could be used to perform the analytical methods according to embodiments of the invention.

B. Detection by Immunoassay

[0162] An immunoassay can be used to detect and analyze markers in a sample. This method comprises: (a) providing an antibody that specifically binds to a marker; (b) contacting a sample with the antibody; and (c) detecting the presence of a complex of the antibody bound to the marker in the sample.

[0163] Methods for producing polyclonal and monoclonal antibodies that react specifically with a cellular marker are known to those of skill in the art. *See, e.g.*, Coligan, *Current Protocols in Immunology* (1991); Harlow & Lane, *Antibodies: A Laboratory Manual* (1988); Goding, *Monoclonal Antibodies: Principles and Practice* (2d ed. 1986); and Kohler & Milstein, *Nature* 256:495-497 (1975). For example, to produce polyclonal antibodies, a purified target protein, is mixed with an adjuvant and used to immunize animals. When high titers of antibody to the target protein are obtained, blood is collected from the animals and antisera are prepared for immunoassays. To produce monoclonal antibodies, spleen cells from an animal immunized with a target protein are immortalized, commonly by fusion with a myeloma cell (*see*, Kohler and Milstein, *Eur. J. Immunol.*, **6**:511-519, 1976). Colonies arising from single immortalized cells are screened for production of antibodies of the desired specificity and affinity for the target protein.

[0164] If the markers are not known proteins in the databases, nucleic acid and amino acid sequences can be determined with knowledge of even a portion of the amino acid sequence of the marker. For example, degenerate probes can be made based on the N-terminal amino acid sequence of the marker. These probes can then be used to screen a genomic or cDNA library created from a sample from which a marker was initially detected. The positive clones can be identified, amplified, and their recombinant DNA sequences can be subcloned using techniques which are well known. *See*, *e.g.*, Ausubel *et al.*, *Current Protocols for Molecular Biology*, 1994 and Sambrook and Russell, *supra*. Based on the polynucleotide sequence encoding a marker, antibodies against the marker can be prepared using any suitable methods known in the art. *See*, *e.g.*, Huse *et al.*, *Science* 246:1275-1281 (1989); Ward *et al.*, *Nature* 341:544-546 (1989).

[0165] After the antibody is provided, a marker can be detected and/or quantified using any of suitable immunological binding assays known in the art (*see*, *e.g.*, U.S. Patent Nos. 4,366,241; 4,376,110; 4,517,288; and 4,837,168). Useful assays include, for example, an enzyme immune assay (EIA) such as enzyme-linked immunosorbent assay (ELISA), a radioimmune assay (RIA), a Western blot assay, or a slot blot assay. These methods are also described in, *e.g.*, *Methods in Cell Biology: Antibodies in Cell Biology*, volume 37 (Asai, ed. 1993); *Basic and Clinical Immunology* (Stites & Terr, eds., 7th ed. 1991); and Harlow & Lane, *supra*.

C. Diagnosis of Metastatic HCC or the Predisposition to Develop HCC

[0166] In another aspect, the present invention provides methods for aiding a diagnosis of the probability of developing metastatic tumors in an HCC patient or a predispositon for developing HCC in a patient with a severe liver disease using one or more markers identified in Tables 2-7. Although valid diagnoses can be made based on as few as one marker selected from the markers in Tables 2-7, it is preferred that multiple markers are used to achieve more reliable results. Preferably, at least 10 cellular markers of Table 2 should be included in the set of markers used to predict an HCC patient's metastatic potential, for example, more preferably at least 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, or 100, and most preferably all 153 markers of Table 2 should be included in the markers used. Similarly, preferably at least 10, more preferably at least 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, or 100, and most preferably all 273 genes of Table 5 should be included in the markers used for determining the risk of developing HCC in a patient with a chronic liver disease. The markers identified in Tables 2-7 can be used alone, in combination with other markers in any of the Tables, or with entirely different markers in aiding in the diagnosis of developing Metastatic HCC or a predisposition for developing HCC by a patient with a severe liver disease. The markers in Tables 2-7 are differentially present in samples of a Metastatic HCC or tissue samples of someone predisposed for HCC relative to a non-metastatic HCC or a subject not predisposed for HCC respectively. For example, some of the markers are expressed at an elevated level and/or are present at a higher frequency in metastatic HCC or tissue samples of someone predisposed for HCC relative to patients with non-metastatic HCC or individuals at low risk for developing HCC. Therefore, detection of one or more of these markers in a person would provide useful information regarding the probability that the person may develop Metastatic HCC or be predisposed to develop HCC.

[0167] Accordingly, embodiments of the invention include methods for aiding in diagnosing the probability of developing Metastatic HCC or in diagnosing the probability of a patient with a severe liver disease developing HCC, wherein the method comprises: (a) detecting at least one marker in a sample, wherein the marker is selected from the markers identified in Tables 2-7; and (b) correlating the detection of the marker or markers with a diagnosis of metastatic HCC or the probability for a liver disease patient to develop HCC. The correlation may take into account the amount of the marker or markers in the sample compared to a control amount of the marker or markers (*e.g.*, a non-metastatic HCC or a subject not predisposed for HCC). The correlation may take into account the presence or absence of the markers in a test sample and the frequency of detection of the same markers in a control. The correlation may take into account both of such factors to facilitate determination of whether a subject has a metastatic HCC or has a sever liver disease that will likely lead to HCC.

[0168] Any suitable samples can be obtained from a subject to detect markers. Preferably, a sample is a liver tissue sample from the subject. If desired, the sample can be prepared as described above to enhance detectability of the markers.

[0169] Any suitable method can be used to detect a marker or markers in a sample. For example, gas phase ion spectrometry or an immunoassay can be used as described above. Using these methods, one or more markers can be detected. Preferably, a sample is tested for the presence of a plurality of markers. Detecting the presence of a plurality of markers, rather than a single marker alone, would provide more information for the diagnostician. Specifically, the detection of a plurality of markers in a sample would increase the percentage of true positive and true negative diagnoses and would decrease the percentage of false positive or false negative diagnoses.

[0170] The detection of the marker or markers is then correlated with a probable diagnosis of developing metastatic HCC or a predispositon for developing HCC by a patient with a severe liver disease. In some embodiments, the detection of the mere presence or absence of a marker, without quantifying the amount of marker, is useful and can be correlated with a probable diagnosis of developing metastatic HCC or a predispositon for developing HCC by a patient with a severe liver disease.

[0171] In other embodiments, the detection of markers can involve quantifying the markers to correlate the detection of markers with a probable diagnosis of developing metastatic HCC or a predispositon for developing HCC by a patient with severe liver disease. For example, increased levels of OPN are observed in patients with metastatic HCC. Thus, if the amount of the markers detected in a subject being tested is higher compared to a control amount, then the subject being tested has a higher probability of developing metastatic HCC or a predispositon for developing HCC by a patient with a severe liver disease.

[0172] When the markers are quantified, it can be compared to a control. A control can be, *e.g.*, the average or median amount of marker present in comparable samples of normal subjects not predisposed to developing metastatic HCC or not predisposed to developing HCC by a patient with severe liver disease. The control amount is measured under the same or substantially similar experimental conditions as in measuring the test amount. For example, if a test sample is obtained from a subject's blood serum sample and a marker is detected using a particular probe, then a control amount of the marker is preferably determined from a serum sample of a patient using the same probe. It is preferred that the control amount of marker is determined based upon a significant number of samples from normal subjects who do not have metastatic HCC or tissue samples of someone not predisposed for HCC so that it reflects variations of the marker amounts in that population.

[0173] Data generated by mass spectrometry can then be analyzed by a computer software. The software can comprise code that converts signal from the mass spectrometer into computer readable form. The software also can include code that applies an algorithm to the analysis of the signal to determine whether the signal represents a "peak" in the signal corresponding to a marker of this invention, or other useful markers. The software also can include code that executes an algorithm that compares signal from a test sample to a typical signal characteristic of "normal" and metastatic HCC or a predispositon for developing HCC by a patient with severe liver disease and determines the closeness of fit between the two signals. The software also can include code indicating which the test sample is closest to, thereby providing a probable diagnosis.

III. Regulation of the Biological Activity of Therapeutic Targets

[0174] Ostoepontin (OPN) and EpCAM have been positively correlated to metastasis in an HCC patient and onset of HCC in a patient with a chronic liver disease, respectively. Therefore, it is one objective of this invention to identify compounds that regulate, particularly inhibit, the activity of OPN or EpCAM.

A. Assays for Biological Functions

[0175] OPN and its alleles and polymorphic variants are secreted phosphoproteins encoded by SEQ ID NO:1 and whose amino acid sequence is disclosed in SEQ ID NO:2. The activity of OPN polypeptides can be assessed using a variety of *in vitro* and *in vivo* assays to determine its functional, chemical, and physical effects, *e.g.*, measuring receptor binding (*e.g.*, radioactive receptor binding), and the like. Further downstream events, such as altered cellular events including cell proliferation, differentiation, etc. may also be used as indirect indicators of modified OPN activity. In addition, such assays can be used to test and screen for antagonists of OPN activity. Antagonists can also be genetically altered versions of OPN, e.g., a dominant negative version of the protein. Such antagonists of OPN activity are useful for treating metastatic HCC.

[0176] The OPN of the assay will be selected from a polypeptide having a sequence of SEQ ID NO: 2 or a conservatively modified variant or fragment thereof. Generally, the amino acid sequence identity will be at least 70%, optionally at least 85%, optionally at least

90-95%. Optionally, the polypeptide of the assays will comprise a domain of OPN, such as a receptor binding domain, an extracellular matrix binding domain, and the like. Either OPN or a domain thereof can be covalently linked to a heterologous protein to create a chimeric protein used in the assays described herein.

[0177] Modulators of OPN activity are tested using OPN polypeptides as described above, either recombinant or naturally occurring. The protein can be isolated, expressed in a cell, secreted from a cell, expressed in tissue or in an animal, either recombinant or naturally occurring. For example, liver slices, dissociated liver cells, or transformed cells can be used.. OPN antagonism is tested using one of the *in vitro* or *in vivo* assays described herein. Furthermore, receptor-binding domains of the OPN protein can be used *in vitro* in soluble or solid state reactions to assay for receptor binding.

[0178] Receptor binding to OPN, a domain, or chimeric protein can be tested in solution, in a bilayer membrane, attached to a solid phase, in a lipid monolayer, or in vesicles. Binding of an antagonist can be tested using, *e.g.*, changes in spectroscopic characteristics (*e.g.*, fluorescence, absorbance, refractive index) hydrodynamic (*e.g.*, shape), chromatographic, or solubility properties.

[0179] Samples or assays that are treated with a potential OPN inhibitor are compared to control samples without the test compound, to examine the extent of antagonism. Control samples (untreated with inhibitors) are assigned a relative OPN activity value of 100. Antagonism of OPN is achieved when the OPN activity value relative to the control is about 90%, optionally 50%, optionally 25-0%.

[0180] Changes in OPN receptor binding may be assessed by determining changes in the ability of the vitronectin receptor to bind OPN in the presence of the antagonist. Generally, the compounds to be tested are present in the range from 1 pM to 100 mM.

[0181] The effects of the test compounds upon the function of the polypeptides can be measured by examining any of the parameters described above. Any suitable physiological change that affects OPN activity can be used to assess the influence of a test compound on the polypeptides of this invention. When the functional consequences are determined using intact cells or animals, one can also measure a variety of effects such as transcriptional changes to both known and uncharacterized genetic markers (*e.g.*, northern blots), changes in cell metabolism such as cell growth or pH changes.

[0182] Similarly, the biological functions of EpCAM may be monitored based on the same general principles and methodologies as described above. For instance, EpCAM is known to play a role in epithelial cell homotypic adhesion, relying on both its extracellular and intracellular domains for proper functioning. Thus, EpCAM's functions can be examined based on, *e.g.*, cell aggregation, specific interactions with its known binding partners (*e.g.*, with actin via its intracellular domain), and disruption of signal transduction it is known to mediate. Various cellular events may serve as indicators of EpCAM activity and to facilitate screening test compounds for EpCAM antagonists.

B. Antagonists

[0183] The compounds tested as antagonists of OPN or EpCAM can be any small chemical compound, or a biological entity, such as a protein, sugar, nucleic acid or lipid. Various antibodies against the proteins are likely candidates for antagonists. For example, many monoclonal antibodies, such as 17-1A and GA733, are known to specifically bind EpCAM and can thus be tested in appropriate assays for their ability to interfere with EpCAM's biological functions.

[0184] Alternatively, antagonists can be genetically altered versions of OPN or EpCAM, such as a so-called "dominant negative" version, a biologically inactive version that suppresses the normal function of its wild type counterpart by competing for limited binding partners. Typically, test compounds will be small chemical molecules and peptides. Essentially any chemical compound can be used as a potential antagonist in the assays of the invention, although most often compounds can be dissolved in aqueous or organic (especially DMSO-based) solutions are used. The assays are designed to screen large chemical libraries by automating the assay steps and providing compounds from any convenient source to assays, which are typically run in parallel (*e.g.*, in microtiter formats on microtiter plates in robotic assays). It will be appreciated that there are many suppliers of chemical compounds, including Sigma (St. Louis, MO), Aldrich (St. Louis, MO), Sigma-Aldrich (St. Louis, MO), Fluka Chemika-Biochemica Analytika (Buchs Switzerland) and the like.

[0185] In one preferred embodiment, high throughput screening methods involve providing a combinatorial chemical or peptide library containing a large number of potential therapeutic compounds (potential modulator or ligand compounds). Such "combinatorial chemical libraries" or "ligand libraries" are then screened in one or more assays, as described herein, to identify those library members (particular chemical species or subclasses) that display a desired characteristic activity. The compounds thus identified can serve as conventional "lead compounds" or can themselves be used as potential or actual therapeutics.

[0186] A combinatorial chemical library is a collection of diverse chemical compounds generated by either chemical synthesis or biological synthesis, by combining a number of chemical "building blocks" such as reagents. For example, a linear combinatorial chemical library such as a polypeptide library is formed by combining a set of chemical building blocks (amino acids) in every possible way for a given compound length (*i.e.*, the number of amino acids in a polypeptide compound). Millions of chemical compounds can be synthesized through such combinatorial mixing of chemical building blocks.

[0187] Preparation and screening of combinatorial chemical libraries is well known to those of skill in the art. Such combinatorial chemical libraries include, but are not limited to, peptide libraries (*see, e.g.*, U.S. Patent 5,010,175; Furka, *Int. J. Pept. Prot. Res.* **37**:487-493, 1991; and Houghton *et al.*, *Nature* **354**:84-88, 1991). Other chemistries for generating chemical diversity libraries can also be used. Such chemistries include, but are not limited to: peptoids (*e.g.*, PCT Publication No. WO 91/19735), encoded peptides (*e.g.*, PCT Publication WO 93/20242), random bio-oligomers (*e.g.*, PCT Publication No. WO 92/00091), benzodiazepines (*e.g.*, U.S. Pat. No. 5,288,514), diversomers such as hydantoins, benzodiazepines and dipeptides (Hobbs *et al.*, *Proc. Nat. Acad. Sci. USA* **90**:6909-6913, 1993), vinylogous polypeptides (Hagihara *et al.*, *J. Amer. Chem. Soc.* **114**:6568, 1992), nonpeptidal peptidomimetics with glucose scaffolding (Hirschmann *et al.*, *J. Amer. Chem. Soc.* **114**:9217-9218, 1992), analogous organic syntheses of small compound libraries (Chen *et al.*, *J. Amer. Chem. Soc.* **116**:2661, 1994), oligocarbamates (Cho *et al.*, *Science* **261**:1303, 1993), and/or peptidyl phosphonates (Campbell *et al.*, *J. Org. Chem.* **59**:658, 1994), nucleic acid libraries (*see* Ausubel, Berger and Sambrook, all *supra*), peptide nucleic acid libraries (*see, e.g.*, U.S. Patent 5,539,083), antibody libraries (*see, e.g.*, Vaughn *et al.*, *Nature Biotechnology*, **14**(3):309-314, 1996 and PCT/US96/10287), carbohydrate libraries (*see, e.g.*, Liang *et al.*, *Science* **274**:1520-1522, 1996 and U.S. Patent 5,593,853), small organic molecule libraries (*see, e.g.*, benzodiazepines, Baum C&EN, Jan 18, page 33, 1993; isoprenoids, U.S. Patent 5,569,588; thiazolidinones and metathiazanones, U.S. Patent 5,549,974; pyrrolidines, U.S. Patents 5,525,735 and 5,519,134; morpholino compounds, U.S. Patent 5,506,337; benzodiazepines, 5,288,514, and the like).

[0188] Devices for the preparation of combinatorial libraries are commercially available (*see, e.g.*, 357 MPS, 390 MPS, Advanced Chem Tech, Louisville KY, Symphony, Rainin, Woburn, MA, 433A Applied Biosystems, Foster City, CA, 9050 Plus, Millipore, Bedford, MA). In addition, numerous combinatorial libraries are themselves commercially available (*see, e.g.*, ComGenex, Princeton, N.J., Tripos, Inc., St. Louis, MO, 3D Pharmaceuticals, Exton, PA, Martek Biosciences, Columbia, MD, etc.).

C. Solid State and soluble high throughput assays

[0189] In one embodiment the invention provide soluble assays using molecules such as a domain such as a receptor binding domain, an extracellular matrix binding domain, etc.; a domain that is covalently linked to a heterologous protein to create a chimeric molecule; OPN or EpCAM; or a cell or tissue expressing OPN or EpCAM, either naturally occurring or recombinant. In another embodiment, the invention provides solid phase based *in vitro* assays in a high throughput format, where the domain, chimeric molecule, OPN or EpCAM, or cell or tissue expressing OPN or EpCAM is attached to a solid phase substrate.

[0190] In the high throughput assays of the invention, it is possible to screen up to several thousand different antagonists or ligands in a single day. In particular, each well of a microtiter plate can be used to run a separate assay against a selected potential modulator, or, if concentration or incubation time effects are to be observed, every 5-10 wells can test a single modulator. Thus, a single standard microtiter plate can assay about 100 (*e.g.*, 96) modulators. If 1536 well plates are used, then a single plate can easily assay from about 100-about 1500 different compounds. It is possible to assay several different plates per day; assay screens for up to about 6,000-20,000 different compounds is possible using the integrated systems of the invention. More recently, microfluidic approaches to reagent manipulation have been developed, *e.g.*, by Caliper Technologies (Palo Alto, CA).

[0191] The molecule of interest can be bound to the solid state component, directly or indirectly, via covalent or non covalent linkage *e.g.*, via a tag. The tag can be any of a variety of components. In general, a molecule which binds the tag (a tag binder) is fixed to a solid support, and the tagged molecule of interest (*e.g.*, the signal transduction molecule of interest) is attached to the solid support by interaction of the tag and the tag binder.

[0192] A number of tags and tag binders can be used, based upon known molecular interactions well described in the literature. For example, where a tag has a natural binder, for example, biotin, protein A, or protein G, it can be used in conjunction with appropriate tag binders (avidin, streptavidin, neutravidin, the Fc region of an immunoglobulin, *etc.*) Antibodies to molecules with natural binders such as biotin are also widely available and appropriate tag binders; *see,* SIGMA Immunochemicals 1998 catalogue SIGMA, St. Louis MO).

[0193] Similarly, any haptenic or antigenic compound can be used in combination with an appropriate antibody to form a tag/tag binder pair. Thousands of specific antibodies are commercially available and many additional antibodies are described in the literature. For example, in one common configuration, the tag is a first antibody and the tag binder is a second antibody which recognizes the first antibody. In addition to antibody-antigen interactions, receptor-ligand interactions are also appropriate as tag and tag-binder pairs. For example, agonists and antagonists of cell membrane receptors (*e.g.*, cell receptor-ligand interactions such as transferrin, c-kit, viral receptor ligands, cytokine receptors, chemokine receptors, interleukin receptors, immunoglobulin receptors and antibodies, the cadherein family, the integrin family, the selectin family, and the like; *see, e.g.*, Pigott & Power, *The Adhesion Molecule Facts Book I* (1993). Similarly, toxins and venoms, viral epitopes, hormones (*e.g.*, opiates, steroids, etc.), intracellular receptors (*e.g.* which mediate the effects of various small ligands, including steroids, thyroid hormone, retinoids and vitamin D; peptides), drugs, lectins, sugars, nucleic acids (linear or cyclic polymer configurations), oligosaccharides, proteins, phospholipids, and antibodies can all interact with various cell receptors.

[0194] Synthetic polymers, such as polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneimines, polyarylene sulfides, polysiloxanes, polyimides, and polyacetates can also form an appropriate tag or tag binder. Many other tag/tag binder pairs are also useful in assay systems described herein, as would be apparent to one of skill upon review of this disclosure.

[0195] Common linkers such as peptides, polyethers, and the like can also serve as tags, and include polypeptide sequences, such as poly gly sequences of between about 5 and 200 amino acids. Such flexible linkers are known to persons of skill in the art. For example, poly(ethelyne glycol) linkers are available from Shearwater Polymers, Inc. Huntsville, Alabama. These linkers optionally have amide linkages, sulfhydryl linkages, or heterofunctional linkages.

[0196] Tag binders are fixed to solid substrates using any of a variety of methods currently available. Solid substrates are commonly derivatized or functionalized by exposing all or a portion of the substrate to a chemical reagent which fixes a chemical group to the surface which is reactive with a portion of the tag binder. For example, groups which are suitable for attachment to a longer chain portion would include amines, hydroxyl, thiol, and carboxyl groups. Aminoalkylsilanes and hydroxyalkylsilanes can be used to functionalize a variety of surfaces, such as glass surfaces. The construction of such solid phase biopolymer arrays is well described in the literature. *See, e.g.*, Merrifield, *J. Am. Chem. Soc.* 85:2149-2154 (1963) (describing solid phase synthesis of, *e.g.*, peptides); Geysen *et al., J. Immun. Meth.* 102:259-274 (1987) (describing synthesis of solid phase components on pins); Frank & Doring, *Tetrahedron* 44:60316040 (1988) (describing synthesis of various peptide sequences on cellulose disks); Fodor *et al., Science*, 251:767-777 (1991); Sheldon *et al., Clinical Chemistry* 39(4):718-719 (1993); and Kozal *et al., Nature Medicine* 2(7):753759 (1996) (all describing arrays of biopolymers fixed to solid substrates). Non-chemical approaches for fixing tag binders to substrates include other common methods, such as heat, cross-linking by UV radiation, and the like.

D. Computer-based assays

[0197] Yet another approach to screen for compounds that modulate OPN or EpCAM activity involves computer assisted drug design, in which a computer system is used to generate a three-dimensional structure of OPN or EpCAM based on the structural information encoded by the amino acid sequence. The input amino acid sequence interacts directly and actively with a pre-established algorithm in a computer program to yield secondary, tertiary, and quaternary structural models of the protein. The models of the protein structure are then examined to identify regions of the structure that have the ability to bind, *e.g.*, ligands. These regions are then used to identify ligands that bind to the protein.

[0198] The three-dimensional structural model of the protein is generated by entering protein amino acid sequences of at least 10 amino acid residues or corresponding nucleic acid sequences encoding an OPN or EpCAM polypeptide into the computer system. For example, the amino acid sequence of an OPN polypeptide or the nucleic acid encoding the polypeptide is selected from the group consisting of SEQ ID NOS:1 or 2, and conservatively modified versions thereof. The amino acid sequence represents the primary sequence or subsequence of the protein, which encodes the structural information of the protein. At least 10 residues of the amino acid sequence (or a nucleotide sequence encoding 10 amino acids) are entered into the computer system from computer keyboards, computer readable substrates that include, but are not limited to, electronic storage media (*e.g.*, magnetic diskettes, tapes, cartridges, and chips), optical media (*e.g.*, CD ROM), information distributed by internet sites, and by RAM. The three-dimensional structural model of the protein is then generated by the interaction of the amino acid sequence and the computer system, using software known to those of skill in the art.

[0199] The amino acid sequence represents a primary structure that encodes the information necessary to form the secondary, tertiary and quaternary structure of the protein of interest. The software looks at certain parameters encoded by the primary sequence to generate the structural model. These parameters are referred to as "energy terms," and primarily include electrostatic potentials, hydrophobic potentials, solvent accessible surfaces, and hydrogen bonding. Secondary energy terms include van der Waals potentials. Biological molecules form the structures that minimize the energy terms in a cumulative fashion. The computer program is therefore using these terms encoded by the primary structure or amino acid sequence to create the secondary structural model.

[0200] The tertiary structure of the protein encoded by the secondary structure is then formed on the basis of the energy terms of the secondary structure. The user at this point can enter additional variables such as whether the protein is membrane bound or soluble, its location in the body, and its cellular location, *e.g.*, cytoplasmic, surface, or nuclear. These variables along with the energy terms of the secondary structure are used to form the model of the tertiary structure. In modeling the tertiary structure, the computer program matches hydrophobic faces of secondary structure with like, and hydrophilic faces of secondary structure with like.

[0201] Once the structure has been generated, potential ligand binding regions are identified by the computer system. Three-dimensional structures for potential ligands are generated by entering amino acid or nucleotide sequences or chemical formulas of compounds, as described above. The three-dimensional structure of the potential ligand is then compared to that of the OPN or EpCAM protein to identify ligands that bind to OPN or EpCAM. Binding affinity between the protein and ligands is determined using energy terms to determine which ligands have an enhanced probability of binding to the protein.

[0202] Computer systems are also used to screen for mutations, polymorphic variants, alleles and interspecies homologs of OPN genes or EpCAM genes. Such mutations can be associated with disease states or genetic traits. As described above, GENECHIP® and related technology can also be used to screen for mutations, polymorphic variants, alleles, and interspecies homologs. Once the variants are identified, diagnostic assays can be used to identify patients having such mutated genes. Identification of the mutated OPN genes, for example, involves receiving input of a first amino acid or nucleic acid sequence encoding OPN, selected from the group consisting of SEQ ID NOS:1 and 2, and conservatively modified versions thereof. The sequence is entered into the computer system as described above. The first nucleic acid or amino acid sequence is then compared to a second nucleic acid or amino acid sequence that has substantial identity to the first sequence. The second sequence is entered into the computer system in the manner described above. Once the first and second sequences are compared, nucleotide or amino acid differences between the sequences are identified. Such sequences can represent allelic differences in OPN genes, and mutations associated with disease states and genetic traits. The same general strategy is also applicable for detecting EpCAM variants and mutants.

E. Kits

[0203] A protein of interest and its homologs are a useful tool for identifying its antagonists. For instance, OPN-specific reagents that specifically hybridize to OPN nucleic acid, such as OPN probes and primers, and OPN specific reagents that specifically bind to the OPN protein, *e.g.*, OPN antibodies are used to examine liver cell expression, signal transduction regulation and diagnose metastatic HCC. The same general methods are applicable to EpCAM as well.

[0204] Nucleic acid assays for the presence and the quantity of OPN or EpCAM polynucleotides in a sample include numerous techniques well known to those skilled in the art, such as Southern blot analysis, northern blot analysis, dot blots, RNase protection, S1 analysis, amplification techniques such as PCR (including RT-PCR) and LCR, and *in situ* hybridization. In *in situ* hybridization, for example, the target nucleic acid, *e.g.*, nucleic acid encoding OPN, is liberated from its cellular surroundings in such as to be available for hybridization within the cell while preserving the cellular morphology for subsequent interpretation and analysis (see Example 1). The following articles provide an overview of the art of *in situ* hybridization: Singer *et al.*, *Biotechniques* 4:230-250 (1986); Haase *et al.*, *Methods in Virology*, vol. VII, pp. 189-226 (1984); and *Nucleic Acid Hybridization: A Practical Approach* (Hames *et al.*, eds. 1987). In addition, OPN or EpCAM protein can be detected with the various immunoassay techniques described above. The test sample is typically compared to both a positive control (*e.g.*, a sample containing recombinant OPN or EpCAM) and a negative control.

[0205] The present invention also provides for kits for screening for modulators of OPN or EpCAM. Such kits can be prepared from readily available materials and reagents. For example, such kits can comprise any one or more of the following materials: OPN (or EpCAM), reaction tubes, and instructions for testing OPN (or EpCAM) activity. Optionally, the kit contains biologically active OPN (or EpCAM). A wide variety of kits and components can be prepared according to the present invention, depending upon the intended user of the kit and the particular needs of the user.

IV. Inhibition of the Expression of Therapeutic Targets

[0206] Another means of inhibiting OPN activity and thereby inhibiting HCC metastasis in an HCC patient is to inhibit OPN expression. Similarly, reduced risk of developing HCC in a patient of a chronic liver disease may be achieved by inhibiting EpCAM expression. A variety of methods well known to those skilled in the art are available for specifically suppressing the expression of a particular gene.

A. Antisense polynucleotides

[0207] Antisense technology has been the most commonly described approach in protocols to achieve gene-specific inactivation and are useful tools in research and diagnostics. For instance, antisense oligonucleotides capable of inhibiting gene expression with high level of specificity are often used by those of ordinary skill in biological sciences to elucidate the function of particular genes.

[0208] The specificity and sensitivity of antisense polynucleotides also make them suitable for therapeutic uses. A large number of U.S. patents and scientific publications relate to the use of antisense oligonucleotides as therapeutic agents in the treatment of diseases in animals and humans. *See, e.g.*, U.S. Patent Nos. 6,080,580; 6,180,403; 6,255,111; 6,306,655; 6,440,739; and 6,524,854. An antisense oligonucleotide contains a sequence complementary to the coding strand of a gene targeted for inactivation (*e.g.*, SEQ ID NO:1 or SEQ ID NO:5) and may be of varying lengths, *e.g.*, from less than 10 nucleotides to more than 100 nucleotides, can be safely and effectively administered to a subject, *e.g.*, a human. An antisense polynucleotide may be an oligomer or a polymer of ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) or mimetics thereof. It may be composed of naturally-occurring nucleobases, sugars and covalent internucleoside (backbone) linkages as well as oligonucleotides having non-naturally-occurring portions that function similarly. Such modified or substituted antisense oligonucleotides are often preferred over native forms because of desirable properties such as, *e.g.*, enhanced cellular uptake, enhanced affinity for nucleic acid target, and increased stability in the presence of nucleases. Antisense oligonucleotides suitable for the present invention may also include oligonucleotides containing modified backbones or non-natural internucleoside linkages. Preferred modified oligonucleotide backbones include, for example, phosphorothioates, chiral phosphorothioates, phosphorodithioates, phosphotriesters, aminoalkylphosphotri-esters, methyl and other alkyl phosphonates including 3'-alkylene phosphonates and chiral phosphonates, phosphinates, phosphoramidates including 3'-amino phosphoramidate and aminoalkylphosphoramidates, thionophosphoramidates, thiono-alkylphosphonates, thionoalkylphosphotriesters, and borano-phosphates having normal 3'-5' linkages, 2'-5' linked analogs of these, and those having inverted polarity wherein the adjacent pairs of nucleoside units are linked 3'-5' to 5'-3' or 2'-5' to 5'-2'. Various salts, mixed salts and free acid forms are also included.

[0209] Furthermore, antisense oligonucleotides suitable for the present invention may correspond to either the coding region or the non-coding region of a target nucleic acid, *e.g.*, OPN or EpCAM.

B. Ribozymes

[0210] The level of mRNA encoded by a gene of interest, *e.g.*, OPN or EpCAM, can also be reduced using ribozymes. Ribozymes are RNA molecules having an enzymatic activity that is capable of cleaving or splicing other separate RNA molecules in a nucleotide sequence specific manner. A ribozyme useful for practicing the present invention is a catalytic or enzymatic RNA molecule with complementarity in a substrate binding region to a specific RNA target, *e.g.*, OPN or EpCAM mRNA, and also has enzymatic activity that is active to cleave and/or splice RNA in that target, thereby inhibiting the expression of the target gene. Methods for designing and using ribozymes to target a particular gene are known to those of skill in the art and described in numerous publications, including U.S. Patent Nos. 6.069,007; 6,107,027; 6,225,291; 6,307,041; 6,482,803; and 6,489,163.

C. Small inhibitory RNA (siRNA)

[0211] Another useful tool to reduce the level of a target mRNA and thus the level of a target protein is small inhibitory RNA (siRNA). siRNA molecules are small double-stranded RNA molecules that elicit a process known as RNA interference, a form of sequence-specific gene inactivation. A proposed mechanism for RNA interference hypothesizes an ATP-dependent cleavage of mRNA molecules activated by a short double-stranded RNA, which is formed between the mRNA and the antisense strand of siRNA. Zamore *et al.*, *Cell* **101**:25-33, 2000. RNA interference has been shown in mammalian cell lines, oocytes, early embryos, and some cell types. *See, e.g.*, Elbashir, Sayda M., *et al.*, *Nature* **411**:494-497, 2001. siRNA coding sequences can be designed based on the sequence of a target gene (*e.g.*, OPN or EpCAM) and inserted into various suitable vectors, such as a plasmid or a viral vector, with properly placed transcription initiation and termination elements. When used in an intended recipient of eukaryotic origin, eukaryotic transcription control elements should be used. The vectors containing siRNA coding sequences can then be delivered to a desired target in accordance with the general methodologies for gene transfer known to those of skill in the art. RNA interference thus provides an alternative means to specifically inhibit the expression of a gene based on its sequence, by causing the rapid degradation of the mRNA of the gene, *e.g.*, OPN or EpCAM.

D. Detection of Reduced Target Gene Expression

[0212] Following the administration of a therapeutic compound containing an agent capable of inhibiting the expression of a target gene, *e.g.*, OPN or EpCAM, the effectiveness of the therapeutic compound can be assessed by comparing the *in vivo* level of the target gene before and after the administration. The general methods for administering a pharmaceutical compound are described in detail in a later section.

[0213] When the inhibition of gene expression is achieved at transcriptional level, i.e., by reduction of the amount of mRNA encoding a target gene, the diminished expression of the target gene may be confirmed using various detection techniques such as Northern blot assays, dot blot, RT-PCR and the like by comparing the mRNA level of the target gene (*e.g.*, OPN or EpCAM) before and after the administration of a therapeutic compound. The general methodologies for performing such analysis are well known to those of ordinary skill in the art and described in various literature (*see, e.g.*, Sambrook and Russell, *supra* and Ausubel et al., *supra*).

[0214] When the inhibition of gene expression is achieved at translational level, i.e., by reduction of the amount of protein encoded by a target gene, the diminished expression of the target gene may be confirmed by comparing the protein level of the target gene (*e.g.*, OPN or EpCAM) before and after the administration of a therapeutic compound using various means of measuring protein levels in tissue samples are well known to the ordinarily skilled artisans. As mentioned above, various immunoassays are routinely used to detect the presence and quantity of a protein of interest, *e.g.*, OPN or EpCAM. A general overview of the applicable technology can be found in Harlow and Lane, *Antibodies, A Laboratory Manual*, 1988.

[0215] Appropriate antibodies for target proteins, e.g., OPN and EpCAM, will be necessary for immunoassays. The general methods for preparing antibodies specific for a target protein are well known in the art and described in an earlier section. Further, some antibodies with desirable specificity may already be available for immunoassays (*e.g.*, various mAb for EpCAM).

[0216] Once antibodies specific for a target protein, *e.g.*, OPN or EpCAM, are available, the level the target protein in a patient can be measured by a variety of immunoassay methods with qualitative and quantitative results available to the clinician. Various samples from the patient, such as blood or liver tissue, can be used in the immunoassays to detected the *in vivo* target protein level according to the general methods described in an earlier section. For a review of immunological and immunoassay procedures in general *see, e.g.*, Stites, *supra;* U.S. Patent Nos. 4,366,241; 4,376,110; 4,517,288; and 4,837,168.

V. Administration of Agents Inhibiting Target Protein Activity and Pharmaceutical Compositions

[0217] Agents that inhibit the activity of a target protein, *e.g.*, OPN or EpCAM, can be administered directly to the human patient for modulation of the target protein activity *in vivo*. Administration is by any of the routes normally used for introducing an antagonist or inhibitor compound into ultimate contact with the tissue to be treated, optionally using the tongue or mouth. The antagonists or inhibitors are administered in any suitable manner, optionally with pharmaceutically acceptable carriers. Suitable methods of administering such antagonists or inhibitors are available and well known to those of skill in the art, and, although more than one route can be used to administer a particular composition, a particular route can often provide a more immediate and more effective reaction than another route.

[0218] Pharmaceutically acceptable carriers are determined in part by the particular composition being administered, as well as by the particular method used to administer the composition. Accordingly, there is a wide variety of suitable formulations of pharmaceutical compositions of the present invention (*see, e.g., Remington's Pharmaceutical Sciences*, 17th ed., 1985).

[0219] The antagonists or inhibitors, alone or in combination with other suitable components, can be made into aerosol formulations (*i.e.*, they can be "nebulized") to be administered via inhalation. Aerosol formulations can be placed into pressurized acceptable propellants, such as dichlorodifluoromethane, propane, nitrogen, and the like.

[0220] Formulations suitable for administration include aqueous and non-aqueous solutions, isotonic sterile solutions, which can contain antioxidants, buffers, bacteriostats, and solutes that render the formulation isotonic, and aqueous and non-aqueous sterile suspensions that can include suspending agents, solubilizers, thickening agents, stabilizers, and preservatives. In the practice of this invention, compositions can be administered, for example, by orally, topically, intravenously, intraperitoneally, intravesically or intrathecally. Optionally, the compositions are administered orally or nasally. The formulations of compounds can be presented in unit-dose or multi-dose sealed containers, such as ampules and vials. Solutions and suspensions can be prepared from sterile powders, granules, and tablets of the kind previously described. The modulators can also be administered as part a of prepared food or drug.

[0221] The dose administered to a patient, in the context of the present invention should be sufficient to effect a beneficial response in the subject over time. The dose will be determined by the efficacy of the particular signal modulators employed and the condition of the subject, as well as the body weight or surface area of the area to be treated. The size of the dose also will be determined by the existence, nature, and extent of any adverse side-effects that accompany the administration of a particular compound or vector in a particular subject.

[0222] In determining the effective amount of an antagonist or inhibitor to be administered in a physician may evaluate circulating plasma levels of the agent, its toxicities, and the production of antibodies against the agent. In general, the dose equivalent of an antagonist or inhibitor is from about 1 ng/kg to 10 mg/kg for a typical subject.

[0223] For administration, antagonists or inhibitors of the present invention can be administered at a rate determined by the LD-50 of the antagonist, and the side-effects of the inhibitor at various concentrations, as applied to the mass and overall health of the subject. Administration can be accomplished via single or divided doses.

VI. Examples

[0224] It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes without limitation.

A. Example 1: Predicting a predisposition for Hepatocellular Carcinoma metastasis

1. MATERIALS AND METHODS

a) Patients and tissue samples.

[0225] All of the HCC samples were obtained with informed consent from patients who underwent curative resection in Liver Cancer Institute, Zhongshan Hospital of Fudan University in China. A total of 107 paired primary HCC, metastatic HCC, and adjacent non-tumor normal liver tissue samples were obtained from 40 patients who were pathologically diagnosed as HCC and underwent hepatectomy at the Liver Cancer Institute, Zhongshan Hospital of Fudan University (formerly Shanghai Medical University) in China. Prior to surgery, each patient was examined by computer tomography of abdomen and chest X-ray, and some patients also were examined by isotope scanning of bone if necessary. Among the 107 paired samples, 81 were from 27 patients who had primary HCC, corresponding adjacent non-tumor liver tissue and metastatic HCC [15 with intra-hepatic spreads (group P) and 12 with tumor thrombus in branch of portal vein (group PT)], and 26 were from 13 patients who had only a single primary HCC and corresponding non-tumor liver tissue (without detectable metastasis at the time of surgery). Tumors and non-tumor tissues were grossly dissected, snap-frozen in liquid nitrogen immediately after removal, and stored at –70°C until use. We confirmed microscopically that tumor tissue samples and their metastases consisted mostly of carcinoma cells and that non-tumor adjacent liver samples did not exhibit any tumor cell invasion. Of the 40 patients, 39 were male, and one was female. Patients' age ranged from

36 years to 74 years, with a median age of 50 years. The size of the primary HCC ranged from 1.3 cm to 17.5 cm in diameter with a median diameter of 7.2 cm, of which 65% (26/40) were > 5 cm in diameter and remaining were ≤5 cm in diameter. Thirty-two cases (80%) had co-existing liver cirrhosis. Serologically, all of the 40 patients with an exception of one were HBV-positive, but no one was HCV-positive. Twenty-seven patients (68%) had an elevated serum concentration of alpha-fetoprotein (AFP) (>20 ng/ml).

b) RNA preparation, cDNA Microarrays and Hybridization.

[0226] Total RNA was extracted from each sample using TRIzol Reagent (Life Technologies, Inc.) according to the manufacturer's specification. The cDNA microarrays were fabricated at the Advanced Technology Center, NCI. Each array contains 9180 cDNA clones with 7102 "named" genes, 1179 EST clones, and 122 Incyte clones. Preparation of fluorescent cDNA targets by a direct labeling approach and the cDNA microarray hybridization were essentially as described by Wu et al., *Oncogene* **20**:3674-3682, 2001. Briefly, the fluorescent targets were prepared as following: 100 μg of total RNA from non-cancerous liver tissue were labeled with Cy3-conjugated deoxynucleotides or 200 μg of total RNA from primary HCC or metastasis were labeled with Cy5-conjugated deoxynucleotides (Amersham) by the oligo dT-primed polymerization using SuperScript II reverse transcriptase (Life Technologies). The targets were then mixed together and added to the microarrays, and then incubated overnight (12-16 hours) at 42°C. Prior to hybridization, each microarray was pre-hybridized at 42°C for at least one hour in pre-hybridization buffer containing 5× SSC, 0.1% SDS and 1% BSA. The slides were washed at room temperature in each with 2x SSC, 0.1% SDS and 1x SSC and 0.2x SSC for 2 min, respectively, and washed in 0.05x SSC for 1 min. Most of samples, when indicated, were done in duplication. The Cy3 and Cy5 fluorescent intensities for each clone were determined by the Axon GenePix 4000 scanner, and were analyzed by the GenePix Pro 3.0 software to subtract the background signals. The expression data were then filtered based on their channel intensities, spots size and flag, and the Cy5/Cy3 ratios were calculated and normalized by median-centering the log-ratio of all genes in each array.

c) Data Analysis and Statistical Analysis.

[0227] Unsupervised hierarchical clustering analysis was done by the CLUSTER and TREEVIEW software using median centered correlation and complete linkage (Eisen et al., *supra*). We also used the BRB-ArrayTools software, an integrated package for the visualization and statistical analysis of cDNA microarray gene expression data developed by the Biometric Research Branch of the National Cancer Institute, for both unsupervised and supervised analyses. The Class Comparison Tool based on univariate F-tests was used to find genes differentially expressed between predefined clinical groups at a significance level of P <0.001 or 0.002. The permutation distribution of the F-statistic, based on 2000 random permutations was also used to confirm statistical significance. In comparing primary to metastatic tumors of the same patient, a paired value t-statistic was used in the same manner. The multi-variate Compound Covariate Predictor (CCP) Tool with a "leave-one-out" cross-validation test using 2000 random permutations at a significant level of P<0.001 was used to classify predefined clinical groups based on their gene expression profiles. In each cross-validation step one sample is omitted and a multivariate CCP is created based on the genes that are univariately significant at the specified level in the training set consisting of the samples not omitted. This CCP is used to classify the omitted sample and it is then noted whether the classification is correct or incorrect. This is repeated with all samples excluded one at a time. The total cross-validated misclassification rate is thereby determined. The statistical significance of the cross-validated misclassification rate is determined by repeating the entire cross-validation procedure to data with the class membership labels randomly permuted 2000 times. The CCP is based on a weighted linear combination of gene expression variables that are univariately significant in the training set with the weights being the corresponding t-statistics as described in Radmacher et al., *supra*. When the CCP was used to classify paired primary and metastatic tissue, the cross-validation was performed with one pair at a time omitted and the classification based on the paired differences in expression for each gene. Averaged gene expression data from duplicated samples were included for the analysis.

[0228] To generate a prediction model to classify HCC with metastasis potential, we randomly selected 10 PN samples and 10 PT samples as a training set. A total of 20-blinded new HCC samples were included as a testing set. The classification of new samples was based on the computation with the following linear combination: $L = \Sigma_i t_i *(x_i - m_i)$, where t_i = t-value for gene i in the classifier, x_i = log-ratio of gene i in the new sample to be classified, and m_i = midpoint between PN and PT groups for gene i (see Table 2). Additional details are available in BRB-ArrayTools Users Guide. The Kaplan-Meier Survival analysis was used to compare patient survival, using an Excel-based WinSTAT software. The statistical P value was generated by the Cox-Mantel log-rank test when PN was compared to P or PT.

d) Semi-quantitative PT-PCR and Western blotting.

[0229] Total RNA was reverse-transcribed with SUPERSCRIPT™ II RNase H$^-$ Reverse Transcriptase and Random hexamers (Invitrogen Inc.). PCR was done with 26 cycles (94°C, 30 sec; 53°C, 30 sec; 72°C, 1 min) followed by an extra cycle at 72°C for 10 min using the following primers: OPN sense 5'-GACTCGAACGACTCTGATGATGTA-3' (SEQ ID NO:3); OPN antisense 5'-CTGGGCAACGGGGATGG-3' (SEQ ID NO:4); and HotStarTaq Master Mix (QIAGEN). QuantumRNA™ 18S (Ambion) was used as an internal standard. Densitometry was used to quantify the amount of OPN, which was normalized by the 18S product. Western blot analysis was done essential as described by Wu et al., *supra*. Briefly, protein lysates from CCL13, SK-Hep-1 and Hep3B cells were prepared in RIPA buffer (50 mM Tris-HCl, pH 7.4/150 mM NaCl/1% Triton X-100/1% deoxycholate/1.0% SDS/1% aprotinin), separated on 10% SDS-PAGE, transferred to an Immobilin-P membrane (Millipore, Bedford, MA), probed with a rat monoclonal anti-OPN antibody (Chemicon International), and visualized by the ECL-based assay (Amersham).

e) Cell lines and In vitro invasion assay.

[0230] Two human hepatoma derived cell lines with different metastatic potential, SK-Hep-1 and Hep3B, and one non-transformed liver cell line, CCL13 (Chang liver cells), were used to determine the functional association of OPN with metastatic potential using the BD BioCoat™ Matrigel™ Invasion Chamber (BD Biosciences) according to the manufacture's instruction. These cells were obtained from American Type Culture Collection. Cells were routinely maintained at 37°C in a humidified atmosphere of 5% CO_2 in EMEM (GIBCOL) medium supplemented with 10% fetal bovine serum, 1× nonessential amino acids, 1× sodium pyruvate, 2 mM glutamine and penicillin/streptomycin. For invasion analysis, cells were plated in the up chamber in serum-free EMEM, and incubated in the absence or presence of either recombinant murine OPN (2 μg/ml) (R&D Systems) or a well-documented neutralizing antibody against OPN (3 μg/ml) (R&D Systems) for 20 hours. The EMEM medium containing 5% FBS was added to the bottom chamber, serving as chemoattractants. The number of cells invading through the Matrigel™ membrane was calculated before and after adding OPN or antibody of OPN for each cell line.

f) Tissue histology analysis.

[0231] Paraffin-embedded tissue blocks were prepared and were subjected to serial sections with a thickness of 5 μm mounted on electrically charged glass slides. Slides were subjected to hematoxylin and eosin (H&E) staining. Two pathologists read these slides independently for the histological diagnosis. For immunohistochemistry analysis, slides were deparafinized and processed for immunostaining as described by Forgues et al., *J. Biol. Chem.* **276**:22797-22803, 2001. Briefly, slides were incubated in microwave oven for 15 min in 1X citrate buffer for antigen retrieval and then quenched with 3% hydrogen peroxide to block the endogenous peroxidase activity for 10 min. Following incubation with 10% donkey serum to block the non-specific binding, the sections were incubated over night at 4EC with a rat monoclonal anti-OPN antibody (Chemicon International). Biotinylated secondary antibodies and streptavidin peroxidase complex (ABC Elite kit, Vector Labs) were used. Chromogenic development was obtained by the immersion of sections in 3-3' di-aminobenzidine (DAB) solution (0.25 mg per ml with 3% hydrogen peroxide). The slides were counter-stained with Harris= Hematoxylin and de-hydrated with alcohol to Xylene, and mounted with Permount (Sigma).

2. RESULTS

a) Metastatic lesions are indistinguishable from their corresponding primary HCC.

[0232] To define the specific changes associated with the metastatic process in HCC, we compared the gene expression profiles of primary HCC samples from individuals with either intra-hepatic spreads (group P) or tumor thrombi in the portal vein (group PT) together with their matched metastatic lesions, i.e., P-M or PT-M, respectively, with their corresponding non-cancerous liver tissues. Initially, we compared the gene expression profiles of 50 primary and metastatic tumor samples from 30 randomly selected individuals [i.e., 10 patients with metastasis-free HCC (group PN), 10 PT patients and 10 P patients]. We attempted to classify them into clinical groups with an unsupervised hierarchical clustering algorithm based on an overall expression similarity profile using either entire 9180 genes or approximately 2487 genes derived from a gene screen filter that excluded genes not significantly more variable than the median at $P<0.01$. However, these clustering approaches did not yield any meaningful classification that corresponded to predefined clinical groups. Similarly, we could not obtain a meaningful classification using 107 genes from filtering genes with an average of 2-fold greater variations in the gene expression ratio when compared with their median. The results of this analysis imply that primary and metastatic HCC differ only by a relatively small subset of genes, whereas the gene clustering algorithm may be dominated by variations among many other genes, therefore, hindering classification.

[0233] To search for such small differences, we applied a supervised class comparison analysis with univariate F-tests and a global permutation test to define genes that were differentially expressed among predefined clinical groups. A comparison of five clinical groups (i.e., P, P-M, PT, PT-M, and PN) yielded a total of 143 significant genes (*P*<0.0005). Multidimensional scaling analysis based on the first three principal components of these 143 significant genes revealed that the PN samples are distinct from the remaining samples, while the P, P-M, PT, and PT-M samples are inseparable (Fig. 1a). Unexpectedly, the gene expression profiles of primary and matched metastatic HCC tumors were not significantly distinguishable.

b) PN is distinct from PT and P.

[0234] To confirm and extend the above findings, we performed a class comparison analysis of 30 primary HCC samples from PN, PT, and P patients. This analysis yielded a total of 383 significant genes (*P*<0.0005). A hierarchical clustering algorithm was then used to sort these 30 PN, P, and PT samples based on the expression profile of these 383 genes (Fig 1b). Two major branches were observed in the hierarchical tree, one associated with PN samples, and the other with P and PT samples. Again, P and PT samples were not fully discriminated (Fig 1b). Thus, primary metastasis-free HCC has a gene expression profile markedly different from that of primary HCC with metastatic lesions in the portal vein or elsewhere in liver parenchyma.

[0235] To further define a gene set that could accurately discriminate into two predefined classes and to identify metastasis-associated genes, we used a supervised machine learning classification algorithm known as compound covariate predictor (CCP), which includes a "leave-one-out" cross-validation test to avoid the statistical problem of over-estimating prediction accuracy that occurs when a model is trained and evaluated with the same samples. This analysis also creates a multivariate predictor for determining which one of the two classes a given sample belongs to, and a gene list that is univariately significant at a given statistically significant level. We divided 50 HCC samples from 30 patients into various pairs based on different clinical criteria and applied the CCP to each pair (Table 1), using an entire gene set with a *P* value < 0.001. At this specified significance level, the expected number of false-positive genes in the classifier is less than 10. The misclassification rate was determined by leave-one-out cross-validation. For each step of the cross-validation in which one sample was left out, the selection of informative genes and the creation of the multi-gene classifier was repeated from scratch. The probability of obtaining as small a cross-validated misclassification rate by chance was obtained by repeating the entire cross-validation procedure using 2000 random permutations of the class labels for the clinical criteria being evaluated. That gave rise to a classifier *P* (Table 1). Using this supervised machine learning classification algorithm, again we found no significant difference between paired PT and PT-M samples (Table 1). Gene expression profiles in P and PT samples were almost identical to their paired metastatic P-M and PT-M samples (Table 1). The number of genes in these classifiers was at the background (false-positive) level. These data are in agreement with the clustering and multidimensional scaling analysis described above.

[0236] In contrast, we accurately predicted primary tumors (100%) from PN and PT samples with a total of 153 significant genes in the classifier (Table 2). The cross-validated misclassification rates were significantly lower than expected by chance (p<0.0005) (Table 1). Similarly, we accurately predicted PN and P samples as well as PN and P/PT samples with significant numbers of genes in the classifiers (Table 1). However, the CCP yielded no statistical significant classification among P, PT, PT-M, and P-M, and the number of genes in these classifiers also was insignificant. Moreover, we found no statistically significant classification when tumor sizes, ages, tumor encapsulation, or cirrhosis were used as clinical categories. These data are consistent with the findings of class comparison analysis including multidimensional scaling and hierarchical clustering algorithm analyses. We conclude that primary and metastatic tumors have a very similar gene expression signature and that primary metastasis-free HCC tumors are distinct from primary HCC tumors with either tumor thrombus in portal vein or intra-hepatic spread.

Table 1. Performance of classifier during "leave-one-out" cross validation *

Classifier category **	Clinical groups	Total number of cases	Number of cases misclassified	Classifier *P* value	Number of genes in the classifiers
PN vs. PT	PN	10	0	<0.0005	153
	PT	10	0		
PN vs. P	PN	10	1	<0.0005	157
	P	10	0		
PN vs. P/PT	PN	10	2	<0.001	256
	P and PT	20	0		

P vs. PT	P	10	3	0.216	20
	PT	10	4		
PT vs. PT-M	paired samples	10	3	0.296	1
P/PT vs. P-M/PT-M	paired samples	20	5	0.132	7
P vs. PT-M	P	10	4	0.248	14
	PT-M	10	3		
PT vs. P-M	PT	10	2	0.163	9
	P-M	10	4		
Tumor sizes	> 5 cm	16	7	0.234	7
	≤ 5 cm	14	4		
Ages	> 45 yr.	17	5	0.334	4
	≤ 45 yr.	13	7		
Tumor encapsulated	presence	9	2	0.037	13
	absence	21	4		
Cirrhosis	presence	14	7	0.798	1
	absence	6	6		

* Compound covariate predictor was used to classify various clinical groups with a total of 9180 gene expression data at a significance level of *P*=0.001. The classifier was based on 2000 random permutations. The expected number of false-positive genes in the classifier is 10.

** PN, single primary HCC; PT, primary HCC with tumor thrombi in portal vein; PT-M, tumor thrombi from paired PT; P, primary HCC with intra-hepatic metastasis; P-M, intra-hepatic metastasis from paired P; P/PT, both P and PT; P-M/PT-M, both P-M and PT-M; tumor sizes, diameter in length.

c) A gene expression-based model from supervised machine learning algorithm can predict HCC patients with metastatic potential.

[0237] The success in distinguishing PN from PT with CCP allowed us to develop a gene-expression-based model to predict HCC patients who had the potential to develop metastasis. We randomly selected primary HCC samples from 10 PN patients and 10 PT patients as a training set to generate a prediction model by "leave-one-out" cross-validated classification. The classification of training samples created a 153-gene list, which provided the base for predicting testing samples, referred to as the "weighted voting" exercise by generating a multi-factorial L value (see Materials and Methods). We included all of the remaining 20 primary HCC samples as a test set (15 P patients, 3 additional PN patients, and 2 additional PT patients). Fig 2 shows the calculated "weighted voting" L value with metastatic samples yielding negative values and non-metastatic samples yielding positive values. All of the test samples with the exception of one "P" sample (S29) were classified to the metastatic group (Fig 2a). Patient follow-up data indicated that one PN patient (S56) was found to develop lung metastases 8 months following surgery, the second PN patient (S57) was cancer-free 9 months after surgery, and the third patient (S55) did not respond to the follow-up request. We also analyzed these samples by multidimensional scaling based on the 153-gene set obtained from the PN/PT comparison. It appears that S29 has a gene expression profile more similar to the P and PT groups than to that of the PN group (Fig 2b), suggesting that S29 should belong to the P and PT groups. Thus, we accurately classified at least 18 of 20 blinded HCC patients (90%) with metastatic potential.

Table 2 153 Significant genes for predicting metastasis and their values necessary for computing multi-factorial L value in the prediction model.

UG cluster	Symbol	Description	t-value	Midpoint	p-value	Unique id
Hs.36566	LIMK1	LIM domain kinase 1	-7.7122	-0.433	0.000000	160082
Hs.75573	CENPE	centromere protein E (312kD)	-7.2301	0.217	0.000001	160128
Hs.81217	FZD2	frizzled (Drosophila) homolog 2	-7.0334	-0.499	0.000002	160028
Hs.146580	ENO2	enolase 2, (gamma, neuronal)	-6.9978	-0.238	0.000002	160068
Hs.222	ITGA9	integrin, alpha 9	-6.699	-0.159	0.000004	160135
Hs.75887	COPA	coatomer protein complex, subunit alpha	-6.4035	-0.241	0.000007	159890
Hs.6727	KIAA0660	Ras-GTPase activating protein SH3 domain	-6.3742	-0.281	0.000007	160103
Hs.89578	GTF2H1	general transcription factor IIH, polypeptide 1	-6.2909	-0.178	0.000006	164987
Hs.180941	VPS41	vacuolar protein sorting 41 (yeast homolog)	-5.9459	-0.331	0.000013	159888
Hs.99236	RGS20	regulator of G-protein signaling 20	-5.8503	-0.264	0.000015	161959
Hs.274	MATK	megakaryocyte-associated tyrosine kinase	-5.8166	-0.366	0.000016	160015
Hs.194816	STOML1	stomatin (EBP72)-like 1	-5.7855	-0.124	0.000018	162695
Hs.79516	BASP1	membrane attached signal protein 1	-5.5974	-0.415	0.000026	159882
Hs.733	EPB42	erythrocyte membrane protein band 4.2	-5.5395	-0.378	0.000029	160067
Hs.87539	ALDH3B2	aldehyde dehydrogenase 3 family, member B2	-5.5356	-0.351	0.000030	166071
Hs.5947	MEL	mel transforming oncogene	-5.434	-0.452	0.000045	160104
Hs.118354	CAT56	CAT56 protein	-5.4077	-0.316	0.000047	165027
Hs.27744	RAB3A	RAB3A, member RAS oncogene family	-5.35	-0.338	0.000044	160099
Hs.7984	PSCD3	pleckstrin homology	-5.3177	-0.143	0.000047	159887
Hs.104519	PLD2	phospholipase D2	-5.2672	-0.275	0.000052	159999
Hs.4748	ADCYAP1R1	adenylate cyclase activating polypeptide 1	-5.2037	-0.166	0.000060	161460
Hs.83155	ALDH3B1	aldehyde dehydrogenase 3 family, member B1	-5.2005	-0.44	0.000088	159838
Hs.283822	RHD	Rhesus blood group, D antigen	-5.1898	-0.369	0.000062	164821
Hs.2175	CSF3R	colony stimulating factor 3 receptor	-5.1684	-0.136	0.000065	160114
Hs.3094	KIAA0063	KIAA0063 gene product	-5.162	-0.325	0.000095	160091
Hs.119273	KIAA0296	KIAA0296 gene product	-5.132	-0.545	0.000070	159951
Hs.23672	LRP6	low density lipoprotein receptor-related protein 6	-5.1081	-1.13	0.000074	162040
Hs.118804	ENO3	enolase 3, (beta, muscle)	-5.0415	-0.76	0.000085	164468

Hs.74502	CTRB1	chymotrypsinogen B1	-5.0381	-0.216	0.000086	159787
Hs.194148	YES1	v-yes-1 Yamaguchi sarcoma viral oncogene	-5.0064	-0.413	0.000092	159875
		Unknown (IncytePD:1404153)	-4.9541	-0.155	0.000103	160122
Hs.772	GYS1	glycogen synthase 1 (muscle)	-4.913	-0.478	0.000112	160222
Hs.153203	MDFI	MyoD family inhibitor	-4.8908	-0.773	0.000138	163880
Hs.247423	ADD2	adducin 2 (beta)	-4.8064	-0.609	0.000141	162687
Hs.22785	GABRE	gamma-aminobutyric acid (GABA) A receptor	-4.8046	-0.188	0.000142	159794
		Unknown (IncytePD:2685601)	-4.7898	-0.307	0.000147	165108
Hs.97087	CD3Z	CD3Z antigen, zeta polypeptide (TiT3 complex)	-4.7723	-0.487	0.000152	160043
Hs.79006	DTYMK	deoxythymidylate kinase (thymidylate kinase)	-4.7693	0.254	0.000153	161858
Hs.26915	SPTBN2	spectrin, beta, non-erythrocytic 2	-4.7666	-0.364	0.000154	160846
		Unknown (IncytePD:2509789)	-4.7523	-0.175	0.000159	164920
Hs.38586	HSD3B1	hydroxy-delta-5-steroid dehydrogenase	-4.7519	-0.392	0.000159	164787
Hs.32966	GUCA2B	guanylate cyclase activator 2B (uroguanylin)	-4.7519	-0.368	0.000159	164851
Hs.12773	ACOX3	acyl-Coenzyme A oxidase 3, pristanoyl	-4.7455	-0.25	0.000187	162487
Hs.2281	CHGB	chromogranin B (secretogranin 1)	-4.7199	-0.269	0.000171	160078
Hs.25197	STUB1	STIP1 homology and U-Box containing protein 1	-4.6897	-0.264	0.000183	160555
Hs.169536	RHAG	Rhesus blood group-associated glycoprotein	-4.6648	-0.326	0.000193	164916
Hs.96	PMAIP1	PMA-induced protein 1	-4.6573	-0.124	0.000196	160112
Hs.153053	CD37	CD37 antigen	-4.6051	-0.652	0.000220	160033
Hs.155227	EPHB4	EphB4	-4.5965	-0.276	0.000257	168938
Hs.92282	PITX2	paired-like homeodomain transcription factor 2	-4.584	-0.149	0.000230	160123
Hs.79123	KIAA0084	KIAA0084 protein	-4.583	-0.296	0.000231	159886
Hs.180878	LPL	lipoprotein lipase	-4.5304	-0.18	0.000259	160485
Hs.75658	PYGB	phosphorylase, glycogen; brain	-4.5152	0.027	0.000268	159778
Hs.286132	MN7	D15F37 (pseudogene)	-4.503	-0.314	0.000275	167399
Hs.57600	AP1S1	adaptor-related protein complex 1	-4.4656	-0.26	0.000299	160042
Hs.67688		ESTs	-4.4472	-0.458	0.000311	162920
Hs.172458	IDS	iduronate 2-sulfatase (Hunter syndrome)	-4.4324	-0.259	0.000322	160243
Hs.80768	CLCN7	chloride channel 7	-4.4298	0.058	0.000324	161279
Hs.347527	SLC20A2	solute carrier family 20, member 2	-4.4173	-0.308	0.000333	159936
Hs.72550	HMMR	hyaluronan-mediated motility receptor (RHAMM)	-4.3918	-0.443	0.000352	167575
		Unknown (IncytePD:1681876)	-4.3868	-0.275	0.000356	166536
Hs.242947	DGKI	diacylglycerol kinase, iota	-4.3835	-0.369	0.000358	161826
Hs.158249	KIAA0406	KIAA0406 gene product	-4.3376	-0.066	0.000397	159825
Hs.182577	INPP5B	inositol polyphosphate-5-phosphatase, 75kD	-4.315	-0.269	0.000417	160074
Hs.37054	EFNA3	ephrin-A3	-4.3085	-0.355	0.000423	161846
Hs.334841	SELENBP1	selenium binding protein 1	-4.3016	-0.481	0.000430	169315
Hs.81454	KHK	ketohexokinase (fructokinase)	-4.2966	-0.36	0.000434	159931
Hs.84790	KIAA0225	KIAA0225 protein	-4.2732	-0.151	0.000582	160472
Hs.94498	LILRA2	leukocyte immunoglobulin-like receptor	-4.2714	-0.308	0.000459	161424
Hs.151393	GCLC	glutamate-cysteine ligase, catalytic subunit	-4.2523	-0.421	0.000479	166059
Hs.151738	MMP9	matrix metalloproteinase 9	-4.2337	-0.473	0.000722	159912
Hs.69707	HCGII-7	HCGII-7 protein	-4.2223	0.802	0.000512	161462
Hs.152251	FZD5	frizzled (Drosophila) homolog 5	-4.2088	-0.386	0.000528	164899
		Unknown (IncytePD:1570216)	-4.2019	-0.336	0.000536	159962
Hs.61712	PDK1	pyruvate dehydrogenase kinase, isoenzyme 1	-4.1746	-0.251	0.000570	160462
Hs.66731	HOXB13	homeo box B13	-4.1722	-0.739	0.000573	159868
Hs.80976	MKI67	antigen identified by monoclonal antibody Ki-67	-4.1699	-0.148	0.000642	160039
Hs.283664	ASPH	aspartate beta-hydroxylase	-4.1693	0.062	0.000576	160084
Hs.76688	CES1	carboxylesterase 1	-4.1577	-1.285	0.000591	164490
Hs.154230	NDP52	nuclear domain 10 protein	-4.1483	-0.178	0.000604	159958
Hs.75596	IL2RB	interleukin 2 receptor, beta	-4.1376	-0.268	0.000688	159942
Hs.4756	FEN1	flap structure-specific endonuclease 1	-4.1222	0.195	0.000640	160035
Hs.673	IL12A	interleukin 12A	-4.0844	-0.082	0.000696	162579
Hs.89230	KCNN3	potassium calcium-activated channel	-4.0745	0.008	0.000711	161095
Hs.799	DTR	diphtheria toxin receptor	-4.0616	-0.421	0.000812	167412
Hs.120360	PLA2G6	phospholipase A2, group VI	-4.0344	-0.577	0.000778	160058
Hs.171075	RFC5	replication factor C (activator 1) 5 (36.5kD)	-4.0263	0.114	0.000792	161332
Hs.99899	TNFSF7	tumor necrosis factor superfamily, member 7	-4.0211	-0.221	0.000801	159817

Hs.9605	CPSF5	cleavage and polyadenylation specific factor 5	-4.0101	0.079	0.000821	159766
Hs.95262	NFRKB	nuclear factor related to kappa B binding protein	-4.0081	-0.162	0.000825	167698
Hs.37129	SCNN1B	sodium channel, nonvoltage-gated 1	-4.0053	-0.244	0.000830	161191
Hs.296371	RAB28	RAB28, member RAS oncogene family	-4.0038	0.343	0.000833	160699
Hs.83795	IRF2	interferon regulatory factor 2	-3.9955	-0.527	0.000848	161188
Hs.85087	LTBP4	latent TGF-beta binding protein 4	-3.9927	-0.34	0.000854	159923
Hs.267448	CGI-85	CGI-85 protein	-3.986	0.219	0.000866	166502
Hs.121521	ABL2	v-abl murine leukemia viral oncogene homolog 2	-3.9746	-0.347	0.000889	166612
Hs.28166	CRSP8	cofactor for Sp1 transcriptional activation	-3.9714	0.07	0.000895	162996
Hs.239706	GAB1	GRB2-associated binding protein 1	-3.9529	-0.347	0.000933	162416
Hs.177687	AKR1C4	aldo-keto reductase family 1, member C4	-3.9499	0.145	0.000939	161753
Hs.25648	TNFRSF5	TNF receptor superfamily, member 5	-3.9371	-0.147	0.000966	166055
Hs.858	RELB	v-rel viral oncogene homolog B	-3.935	-0.12	0.000971	164810
Hs.155314	KIAA0095	KIAA0095 gene product	-3.9244	-0.206	0.000994	162213
Hs.8358	FLJ20366	hypothetical protein FLJ20366	3.9437	0.201	0.000952	164145
Hs.112819		ESTs	3.9573	0.217	0.000924	168969
Hs.126263		ESTs, Highly similar to A38712 fibrillarin	3.9651	0.925	0.000908	167474
Hs.10669	DDEF1	development and differentiation enhancing factor 1	3.9709	-0.062	0.000896	164026
Hs.99216		ESTs, similar to ALU8	3.9802	0.288	0.000878	169148
Hs.98738	GRTH	gonadotropin-regulated testicular RNA helicase	3.9911	-0.198	0.000857	166657
Hs.28274		Homo sapiens cDNA: FLJ22049 fis	3.9912	0.208	0.000857	163989
Hs.186564		ESTs	4.0128	0.177	0.000816	163409
Hs.34045	FLJ20764	hypothetical protein FLJ20764	4.0142	0.325	0.000814	168581
Hs.3686	KIAA0978	KIAA0978 protein	4.0211	0.308	0.000801	164187
Hs.172148		ESTs	4.0307	0.179	0.000784	163746
Hs.239499	KIAA0185	KIAA0185 protein	4.0679	0.17	0.000722	168413
Hs.169341	HTPAP	HTPAP protein	4.1104	0.608	0.000657	163274
Hs.44131	KIAA0974	KIAA0974 protein	4.1179	0.828	0.000646	164589
Hs.2969	SKI	v-ski avian sarcoma viral oncogene homolog	4.1484	0.323	0.000604	164039
Hs.80618	FLJ20015	hypothetical protein	4.1716	0.258	0.000573	163363
Hs.136309	SH3GLB1	SH3-domain, GRB2-like, endophilin B1	4.1832	0.339	0.000559	162621
Hs.274293		Homo sapiens mRNA; cDNA DKFZp761G1111	4.1964	-0.013	0.000543	165504
Hs.21479	UBN1	ubinuclein 1	4.2096	0.554	0.000527	167995
Hs.155160	SRP46	Splicing factor, arginine/serine-rich, 46kD	4.2889	0.291	0.000442	168577
Hs.105584	RPS6KA4	ribosomal protein S6 kinase, 90kD, polypeptide 4	4.3239	0.349	0.000409	168189
Hs.279886	RANBP9	RAN binding protein 9	4.336	0.365	0.000398	168730
Hs.197298	NS1-BP	NS1-binding protein	4.346	0.593	0.000389	168257
		Unknown (IncytePD:2895226)	4.3857	-0.2	0.000357	161881
Hs.36793	FLJ23188	hypothetical protein FLJ23188	4.3907	0.454	0.000353	168869
Hs.17384		ESTs	4.3978	-0.04	0.000347	163225
Hs.78524	HTCD37	TcD37 homolog	4.4097	0.381	0.000338	167570
Hs.2301	DBH	dopamine beta-hydroxylase	4.4196	0.743	0.000375	168202
Hs.118795	FLJ10008	hypothetical protein FLJ10008	4.4386	-0.064	0.000317	166653
Hs.33074		Homo sapiens, clone IMAGE:3606519	4.5036	0.135	0.000275	168589
Hs.4988		Homo sapiens clone 24711 mRNA sequence	4.5042	0.016	0.000274	160165
Hs.288872	FLJ21439	hypothetical protein FLJ21439	4.5242	0.29	0.000263	168393
Hs.323712	KIAA0615	KIAA0615 gene product	4.5292	0.024	0.000260	163625
Hs.14051		Homo sapiens mRNA; cDNA DKFZp434A2417	4.5538	0.215	0.000246	168381
Hs.296287		Similar to bromodomain-containing 4	4.5576	0.499	0.000244	169290
Hs.57847		ESTs, similar to CASPASE-4 PRECURSOR	4.63	0.264	0.000208	165194
Hs.26289		ESTs	4.7062	0.948	0.000176	169360
Hs.11123	DKFZP564G092	DKFZP564G092 protein	4.9593	0.476	0.000101	163064
Hs.288908		cDNA: FLJ21913 fis, clone HEP03888	4.9597	0.556	0.000101	168395
Hs.77495	UBXD2	UBX domain-containing 2	4.9758	0.676	0.000098	160190
Hs.24341	TAZ	transcriptional co-activator with PDZ-binding motif	5.0014	0.127	0.000093	164176
Hs.50133		ESTs	5.153	0.243	0.000067	168567
Hs.262958	DKFZP434B044	hypothetical protein DKFZp434B044	5.1851	0.378	0.000075	169042
Hs.53478		Homo sapiens cDNA FLJ12366 fis	5.2202	0.111	0.000058	168383
Hs.80658	UCP2	uncoupling protein 2	5.2483	1.308	0.000054	168158

Hs.209065	FLJ14225	hypothetical protein FLJ14225	5.3394	0.468	0.000045	164339
Hs.92357	GALK1	galactokinase 1	5.6456	1.15	0.000037	169675
Hs.50373		ESTs	5.7625	0.94	0.000029	165500
Hs.266959	HBG1	hemoglobin, gamma A	5.9704	1.164	0.000026	168326
Hs.25566		ESTs	6.1164	0.182	0.000009	168197
Hs.25277	FLJ21065	hypothetical protein FLJ21065	6.1957	0.116	0.000008	164202

[0238] The above outcome predictor separated 40 patients into two groups, one being metastatic and the other being non-metastatic. Kaplan-Meier survival data indicates that patients who were predicted to be metastatic had significantly shortened survival when compared with patients without detectable metastasis (Fig 2c). Because the mortality of HCC patients relies largely on whether they develop intra-hepatic metastasis, our results indicate that the gene set used in the classifier provides an accurate gene expression signature reflecting liver cancer metastasis and survival.

d) Osteopontin promotes HCC metastasis.

[0239] The above study indicates that the genes necessary for intra-hepatic metastasis should be included in the prediction model. However, the list of 153 genes from the prediction model was based on a stringent criterion (*P* value at 0.001) to minimize the number of false-positive genes in the classifier that is needed for an accurate classification. Such stringent criterion may exclude many genes that could be significant for metastasis progression. To broaden our search, we performed univariate F-tests with a total of 2000 random permutations at a *P* value of < 0.002 on 10 PN and 10 PT primary HCC samples. This analysis yielded a total of 224 significant genes with less than 20 expected false-positives (see Table 3). To identify genes that may contribute to liver cancer metastasis, we inspected the 224-gene list and sorted the top 30 genes whose expressions were altered largely in PT and PT-M, but rarely in PN (see Table 4). These genes were median-centered and visualized by hierarchical clustering algorithm using centered correlation and complete linkage (Fig. 3a).

[0240] A gene with an average of over 3-fold overexpression in PT, but not in PN, was identified as osteopontin (OPN) (SEQ ID NO:1), a secreted phosphoprotein that has recently been found to be highly expressed in metastatic breast tumors as well as malignant lung, colon, and prostate cancers. Comparison of microarray expression data indicated that OPN expression is elevated in most PT samples and their corresponding PT-M samples, but to a much lesser degree in the PN samples (Fig 3b). OPN overexpression in PT samples, but not in PN samples, was confirmed by a semi-quantitative RT-PCR analysis (Fig 3c and d). Immunohistochemical analysis (IHC) of OPN was also performed on 29 primary HCC (including 16 new HCC cases) and 8 normal livers from healthy organ donors. The immunoreactivity of OPN on these samples was evaluated by a blinded fashion. Only metastatic tumors were positive for cytoplasmic OPN staining, especially in the area with high density of vasculature (Fig. 4). The IHC results mostly agreed with microarray and RT-PCR data (61% positive cases; 11 of 18 metastatic HCC) (data not shown). Taken together, these studies demonstrate a good diagnostic value of OPN for metastatic HCC patients.

[0241] To determine the role of OPN in metastasis, we compared the level of OPN in human HCC cell lines by Western blot and *in vitro* invasiveness by Matrigel assay. The level of OPN was high in SK-Hep-1, intermediate in Hep3B and low in CCL13 (Fig 5a), which coincided with their invasiveness (Fig 5b). An OPN neutralizing antibody significantly blocked invasion of SKHep-1 ($p<0.001$) and Hep3B cells ($p<0.04$). However, recombinant murine OPN did not show any statistically significant stimulation ($p>0.05$) on Hep3B and Sk-Hep-1 cells, implying that either OPN produced by tumor cells is sufficient for maintaining an invasive phenotype, or that lesser effect is due to species difference. Similar results were obtained with 5 additional HCC cell lines (Fig 5c). However, the neutralizing antibody had little effect on cell viability and migration (Fig 5c, right panel).

[0242] To extend above finding, we examined the role of OPN on pulmonary metastasis of HCC cells in nude mice. HCCLM3 cell line is a clone derived from MHCC97 cells with a high degree of pulmonary metastasis following subcutaneous (s.c.) injection (Li et al., *J. Cancer Res. Clin. Oncology*, 2002). Consistent with our recent data, a 100% of tumorigenicity was achieved in 1 week after s.c. injection. There was no significant difference in the size of primary tumors between control and anti-OPN groups (Figure 5 E), which is consistent with our *in vitro* results that anti-OPN does not affect HCC cell growth. At the 5th week, pulmonary metastatic lesions were detected in every mouse in the control group with most of the grade I-II tumor clusters and some grade III-IV tumor clusters (Figure 5 E, F). The control mice had an average of 11.1 ± 2.9 tumor clusters per lung. In contrast, only about a half of mice in the anti-OPN group had developed lung metastasis and remaining mice developed mostly grade I tumor clusters with a combined average of 2.6 ± 1.0 tumor clusters per lung, and this effect was statistically significant ($p<0.01$). Therefore, anti-OPN antibody shows a significant inhibitory effect on the lung metastasis of HCCLM3 cells.

Table 3. 224 Significant genes for predicting metastasis and their values necessary for computing multifactorial L value in the prediction model.

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.313	OPN	Osteopontin	1.07	3.29	0.00122	4	161923	IncytePD:4327691
Hs.69707	HCGII-7	HCGII-7 protein	1.07	2.85	0.000512	6	161462	IncytePD:1656490
Hs.177687	AKR1C4	aldo-keto reductase family 1, member C4	0.58	2.11	0.000939	10p15-p14	161753	IncytePD:5033671
		Unknown	0.82	1.74	0.0018		161371	IncytePD:3421817
Hs.276916	NR1D1	nuclear receptor subfamily 1, group D, member 1	0.74	1.71	0.00181	17q11.2	166707	IncytePD:1904760
Hs.211569	GPRK5	G protein-coupled receptor kinase 5	0.99	1.69	0.00147	10q24-qter	161133	IncytePD:1418741
Hs.75573	CENPE	centromere protein E (312kD)	0.82	1.65	1.00E-06	4q24-q25	160128	IncytePD:3081067
Hs.283664	ASPH	aspartate beta-hydroxylase	0.7	1.56	0.000576	8q12.1	160084	IncytePD:3693273
Hs.296371	RAB28	RAB28, member RAS oncogene family	1.07	1.5	0.000833	4p16.1	160699	IncytePD:1457948
Hs.89267		ESTs	2.49	1.48	0.00132	1	163570	IncytePD:1633393
Hs.79411	RPA2	replication protein A2 (32kD)	1.02	1.47	0.00135	1p35	167684	IncytePD:1729876
Hs.79006	DTYMK	deoxythymidylate kinase (thymidylate kinase)	0.98	1.45	0.000153	2	161858	IncytePD:4818795
Hs.26289		ESTs	2.59	1.44	0.000176	17	169360	IncytePD:674211
Hs.4756	FEN1	flap structure-specific endonuclease 1	0.91	1.44	0.00064	11q12	160035	IncytePD:2050085
Hs.44131	KIAA0974	KIAA0974 protein	2.19	1.44	0.000646	10	164589	IncytePD:4540
Hs.267448	CGI-85	CGI-85 protein	0.96	1.42	0.000866	11q13	166502	IncytePD:2603232
Hs.171075	RFC5	replication factor C (activator 1) 5 (36.5kD)	0.83	1.41	0.000792	12q24.2-q24.3	161332	IncytePD:3590056
Hs.77495	UBXD2	UBX domain-containing 2	1.88	1.36	9.78E-05	2p14-q21.3	160190	IncytePD:1940994
Hs.184175	C2orf3	chromosome 2 open reading frame 3	0.84	1.36	0.00139	2p11.2-p11.1	166136	IncytePD:2779394
Hs.146580	ENO2	enolase 2, (gamma, neuronal)	0.55	1.31	1.56E-06	12p13	160068	IncytePD:1672630
Hs.96	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1	0.64	1.31	0.000196	18q22	160112	IncytePD:1931117
Hs.80768	CLCN7	chloride channel 7	0.83	1.3	0.000323	16p13	161279	IncytePD:1522646

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
		Unknown	0.8	1.3	0.00122		165687	IncytePD:404768
Hs.9605	CPSF5	cleavage and polyadenylation specific factor 5, 25 kD subunit	0.87	1.29	0.000821	16	159766	IncytePD:1813371
Hs.20295	CHEK1	CHK1 (checkpoint, S.pombe) homolog	0.85	1.28	0.00185	11q24-q24	161544	IncytePD:2594058
Hs.37288	NR1D2	nuclear receptor subfamily 1, group D, member 2	0.66	1.27	0.00168	3	159975	IncytePD:2643094
Hs.75658	PYGB	phosphorylase, glycogen; brain	0.83	1.26	0.000268	20p11.2-p11.1	159778	IncytePD:1975552
Hs.32058	C1orf19	chromosome 1 open reading frame 19	1.85	1.26	0.00154	1q25	169022	IncytePD:2285569
Hs.24994	LOC51098	CGI-53 protein	1.75	1.25	0.0018	20	166179	IncytePD:2347842
Hs.11123	DKFZP564G09 2	DKFZP564G092 protein	1.56	1.24	0.000101	10cen-q26.11	163064	IncytePD:2071705
Hs.13421	KIAA0056	KIAA0056 protein	0.9	1.24	0.00119	11	159874	IncytePD:1561606
Hs.28166	CRSP8	cofactor required for Sp1 transcriptional activation, subunit 8 (34kD)	0.9	1.23	0.000895	5	162996	IncytePD:1283515
Hs.57973	CARD10	caspase recruitment domain protein 10	1.7	1.23	0.00197	22q13.1	165430	IncytePD:3739467
Hs.274313	IGFBP6	insulin-like growth factor binding protein 6	0.87	1.21	0.00192	12q13	160319	IncytePD:1968126
Hs.209065	FLJ14225	hypothetical protein FLJ14225	1.62	1.18	4.48E-05	1q21	164339	IncytePD:1486385
Hs.34526	TYMSTR	G protein-coupled receptor	0.89	1.18	0.00101	3p21	161635	IncytePD:2610374
Hs.80658	UCP2	uncoupling protein 2 (mitochondrial, proton carrier)	5.23	1.17	5.44E-05	11q13	168158	IncytePD:1907952
Hs.197298	NS1-BP	NS1-binding protein	1.95	1.17	0.000389	1q25.1-q31.1	168257	IncytePD:630045
Hs.222	ITGA9	integrin, alpha 9	0.69	1.16	3.74E-06	3p21.3	160135	IncytePD:2487318
Hs.288908		Homo sapiens cDNA: FLJ21913 fis, clone HEP03888	1.87	1.16	0.000101		168395	IncytePD:1938947
Hs.21479	UBN1	ubinuclein 1	1.86	1.16	0.000527	16p13.3	167995	IncytePD:1541201
Hs.152981	CDS1	CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 1	0.81	1.16	0.0011	4q21	165060	IncytePD:1406071
		Unknown	0.68	1.15	0.000159		164920	IncytePD:2509789
Hs.155223	STC2	stanniocalcin 2	0.88	1.15	0.00122	5p14.2-q15	160310	IncytePD:2823476
Hs.1309	CD1A	CD1A antigen, a polypeptide	0.79	1.15	0.00161	1q22-q23	165058	IncytePD:2906655
Hs.89230	KCNN3	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	0.89	1.14	0.000711	1q21.3	161095	IncytePD:1747441

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.331328	FLJ13213	hypothetical protein FLJ13213	0.7	1.14	0.00136	15	166434	IncytePD:2382190
		Unknown	0.72	1.13	0.000103		160122	IncytePD:1404153
Hs.169341	HTPAP	HTPAP protein	2.05	1.13	0.000657	8	163274	IncytePD:2626340
Hs.78524	HTCD37	TcD37 homolog	1.51	1.12	0.000338	1q21	167570	IncytePD:1430538
Hs.36793	FLJ23188	hypothetical protein FLJ23188	1.68	1.12	0.000353	3p13-q13.33	168869	IncytePD:2669866
Hs.154230	NDP52	nuclear domain 10 protein	0.7	1.12	0.000604	17q21.3	159958	IncytePD:1818836
Hs.25648	TNFRSF5	tumor necrosis factor receptor superfamily, member 5	0.83	1.12	0.00132	20q12-q13.2	160900	IncytePD:1638346
Hs.6727	KIAA0660	Ras-GTPase activating protein SH3 domain-binding protein 2	0.61	1.11	6.92E-06	4q21.1-q21.3	160103	IncytePD:1899625
Hs.8402	ADCY3	adenylate cyclase 3	0.77	1.11	0.00128	2p24-p22	167084	IncytePD:1966824
Hs.279886	RANBP9	RAN binding protein 9	1.52	1.1	0.000398	6p23	168730	IncytePD:1781729
Hs.66718	RAD54L	RAD54 (S.cerevisiae)-like	0.91	1.1	0.00103	1p32	166204	IncytePD:2645840
Hs.10095	LOC56930	hypothetical protein from EUROIMAGE 1669387	1.61	1.1	0.00129	19p13.3	168579	IncytePD:322585
Hs.19348	FLJ13119	hypothetical protein FLJ13119	1.56	1.1	0.00131	15	169102	IncytePD:1978282
Hs.194816	STOML1	stomatin (EBP72)-like 1	0.77	1.09	1.75E-05	15q24-q25	162695	IncytePD:1741526
		Unknown	0.6	1.09	0.000147		165108	IncytePD:2685601
Hs.84790	KIAA0225	KIAA0225 protein	0.75	1.09	0.000582	7	160472	IncytePD:482519
Hs.80976	MKI67	antigen identified by monoclonal antibody Ki-67	0.75	1.09	0.000642	10q25-qter	160039	IncytePD:2470485
Hs.89578	GTF2H1	general transcription factor IIH, polypeptide 1 (62kD subunit)	0.72	1.08	6.25E-06	11p15.1-p14	164987	IncytePD:37249
Hs.27744	RAB3A	RAB3A, member RAS oncogene family	0.58	1.08	4.38E-05	19p13.2	160099	IncytePD:1381611
Hs.2281	CHGB	chromogranin B (secretogranin 1)	0.64	1.08	0.000171	20pter-p12	160078	IncytePD:2821341
Hs.92282	PITX2	paired-like homeodomain transcription factor 2	0.75	1.08	0.00023	4q25-q27	160123	IncytePD:2794019
Hs.194694	MAP3K6	mitogen-activated protein kinase kinase kinase 6	0.8	1.08	0.00119	1	161091	IncytePD:1650939
Hs.7984	PSCD3	pleckstrin homology, Sec7 and coiled/coil domains 3	0.77	1.07	4.69E-05	7	159887	IncytePD:3029341
Hs.158249	KIAA0406	KIAA0406 gene product	0.85	1.07	0.000397	20	159825	IncytePD:1618693
Hs.61712	PDK1	pyruvate dehydrogenase kinase, isoenzyme 1	0.66	1.07	0.00057	2p14-q14.3	160462	IncytePD:268900

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.126263		ESTs, Highly similar to A38712 fibrillarin	3.36	1.07	0.000908	5	167474	IncytePD:1266194
Hs.25648	TNFRSF5	tumor necrosis factor receptor superfamily, member 5	0.76	1.07	0.000966	20q12-q13.2	166055	IncytePD:549096
Hs.239818	PIK3CB	phosphoinositide-3-kinase, catalytic, beta polypeptide	0.78	1.07	0.00114	3q24	160414	IncytePD:267803
Hs.656	CDC25C	cell division cycle 25C	0.79	1.07	0.00118	5q31	165792	IncytePD:876382
Hs.288319	SART1	squamous cell carcinoma antigen recognized by T cells	0.62	1.07	0.00164	11cen-q12.3	164720	IncytePD:2205225
Hs.180878	LPL	lipoprotein lipase	0.73	1.06	0.000259	8p22	160485	IncytePD:647128
Hs.136309	SH3GLB1	SH3-domain, GRB2-like, endophilin B1	1.51	1.06	0.000559	1p22	162621	IncytePD:1552337
Hs.3686	KIAA0978	KIAA0978 protein	1.44	1.06	0.000801	20	164187	IncytePD:2234421
Hs.146007		Homo sapiens clone IMAGE 21721	1.75	1.06	0.00173	2	162822	IncytePD:3143449
Hs.22785	GABRE	gamma-aminobutyric acid (GABA) A receptor, epsilon	0.73	1.05	0.000142	Xq28	159794	IncytePD:3213034
Hs.296287		Similar to bromodomain-containing 4, clone IMAGE:3542455	1.9	1.05	0.000244		169290	IncytePD:2310314
Hs.152251	FZD5	frizzled (Drosophila) homolog 5	0.56	1.05	0.000528	2q33-q34	164899	IncytePD:3129290
Hs.673	IL12A	interleukin 12A (natural killer cell stimulatory factor 1)	0.85	1.05	0.000696	3p12-q13.2	162579	IncytePD:2760318
Hs.155160	SRP46	Splicing factor, arginine/serine-rich, 46kD	1.43	1.04	0.000442	11q22	168577	IncytePD:886075
Hs.151393	GCLC	glutamate-cysteine ligase, catalytic subunit	0.53	1.04	0.000479	6p12	166059	IncytePD:818192
Hs.82927	AMPD2	adenosine monophosphate deaminase 2 (isoform L)	0.82	1.04	0.00163	1p13.3	162188	IncytePD:1968035
Hs.2175	CSF3R	colony stimulating factor 3 receptor (granulocyte)	0.8	1.03	6.46E-05	1p35-p34.3	160114	IncytePD:1596060
Hs.286132	MN7	D15F37 (pseudogene)	0.63	1.03	0.000275	15q11-q13	167399	IncytePD:2739109
Hs.5716	KIAA0310	KIAA0310 gene product	1.39	1.03	0.00185	9q34.2-9q34.3	169169	IncytePD:1880859
Hs.104519	PLD2	phospholipase D2	0.67	1.02	5.23E-05	17p13.1	159999	IncytePD:3472725
Hs.74502	CTRB1	chymotrypsinogen B1	0.73	1.02	8.55E-05	16q23-q24.1	159787	IncytePD:2070278

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.288872	FLJ21439	hypothetical protein FLJ21439	1.46	1.02	0.000263	15q14	168393	IncytePD:1998519
Hs.57600	AP1S1	adaptor-related protein complex 1, sigma 1 subunit	0.69	1.02	0.000299	7	160042	IncytePD:1804181
Hs.17409	CRIP1	cysteine-rich protein 1 (intestinal)	1.5	1.02	0.00123	7q11.23	169514	IncytePD:2121863
Hs.4748	ADCYAP1R1	adenylate cyclase activating polypeptide 1 (pituitary) receptor type I	0.79	1.01	5.99E-05	7p14	161460	IncytePD:3214293
Hs.25197	STUB1	STIP1 homology and U-Box containing protein 1	0.69	1.01	0.000183	16	160555	IncytePD:1315677
Hs.34045	FLJ20764	hypothetical protein FLJ20764	1.56	1.01	0.000814	14	168581	IncytePD:901577
Hs.95262	NFRKB	nuclear factor related to kappa B binding protein	0.79	1.01	0.000825	11q24-q25	167698	IncytePD:1685182
Hs.858	RELB	v-rel avian reticuloendotheliosis viral oncogene homolog B (nuclear factor of kappa light polypeptide gene enhancer in B-cells 3)	0.84	1.01	0.000971	19q13.2	164810	IncytePD:1859449
Hs.180941	VPS41	vacuolar protein sorting 41 (yeast homolog)	0.63	1	1.26E-05	7p14-p13	159888	IncytePD:2910949
Hs.80618	FLJ20015	hypothetical protein	1.43	1	0.000573	17q25	163363	IncytePD:2043391
Hs.75596	IL2RB	interleukin 2 receptor, beta	0.69	1	0.000688	22q13.1	159942	IncytePD:3936210
Hs.99216		ESTs, similar to ALU8_HUMAN ALU SUBFAMILY SX SEQUENCE	1.49	1	0.000878	15	169148	IncytePD:2285350
Hs.155314	KIAA0095	KIAA0095 gene product	0.75	1	0.000994	16q22.1-q22.3	162213	IncytePD:268942
Hs.687	CYP4B1	cytochrome P450, subfamily IVB, polypeptide 1	0.85	1	0.00114	1p34-p12	167183	IncytePD:856900
Hs.75807	PDLIM1	PDZ and LIM domain 1 (elfin)	1.63	1	0.00145	10q22-q26.3	160215	IncytePD:2132217
		Unknown	0.66	1	0.00164		159927	IncytePD:2606307
Hs.41587	RAD50	RAD50 (S. cerevisiae) homolog	0.57	1	0.00183	5q31	160088	IncytePD:1515426
Hs.75887	COPA	coatomer protein complex, subunit alpha	0.73	0.99	6.55E-06	1q23-q25	159890	IncytePD:3296228
Hs.25566		ESTs	1.3	0.99	8.89E-06	22	168197	IncytePD:948796
Hs.274	MATK	megakaryocyte-associated tyrosine kinase	0.61	0.99	1.64E-05	19p13.3	160015	IncytePD:1515980
Hs.347527	SLC20A2	solute carrier family 20 (phosphate transporter), member 2	0.66	0.99	0.000333	8p12-q21	159936	IncytePD:2942938
Hs.242947	DGKI	diacylglycerol kinase, iota	0.61	0.99	0.000358	7q32.3-q33	161826	IncytePD:3108609
Hs.2301	DBH	dopamine beta-hydroxylase (dopamine beta-monooxygenase)	2.82	0.99	0.000375	9q34	168202	IncytePD:1294466

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.172148		ESTs	1.29	0.99	0.000784	5	163746	IncytePD:929090
Hs.99899	TNFSF7	tumor necrosis factor (ligand) superfamily, member 7	0.74	0.99	0.000801	19p13	159817	IncytePD:2017463
Hs.99236	RGS20	regulator of G-protein signaling 20	0.71	0.98	1.53E-05	8	161959	IncytePD:4711030
Hs.262958	DKFZP434B044	hypothetical protein DKFZp434B044	1.72	0.98	7.45E-05	16	169042	IncytePD:211389
Hs.57847		ESTs, similar to ICE4_HUMAN CASPASE-4 PRECURSOR	1.47	0.98	0.000208	11	165194	IncytePD:1362601
Hs.155227	EPHB4	EphB4	0.7	0.98	0.000257	7q22	168938	IncytePD:2056923
Hs.72550	HMMR	hyaluronan-mediated motility receptor (RHAMM)	0.55	0.98	0.000352	5q33.2-qter	167575	IncytePD:3622417
		Unknown	0.7	0.98	0.000356		166536	IncytePD:1681876
		Unknown	0.64	0.98	0.000536		159962	IncytePD:1570216
Hs.296348	DLST	dihydrolipoamide S-succinyltransferase	0.49	0.98	0.00151	14q24.3	165547	IncytePD:1830335
Hs.3094	KIAA0063	KIAA0063 gene product	0.66	0.97	9.45E-05	22q13.1	160091	IncytePD:3227603
Hs.32966	GUCA2B	guanylate cyclase activator 2B (uroguanylin)	0.62	0.97	0.000159	1p34-p33	164851	IncytePD:1806219
		Unknown	0.79	0.97	0.00121		164791	IncytePD:3190386
Hs.190189		ESTs	1.33	0.97	0.00172	1	163286	IncytePD:1679304
Hs.733	EPB42	erythrocyte membrane protein band 4.2	0.62	0.96	2.93E-05	15q15-q21	160067	IncytePD:2052032
Hs.5947	MEL	mel transforming oncogene- RAB8 homolog	0.56	0.96	4.47E-05	19p13.1	160104	IncytePD:1553995
Hs.169536	RHAG	Rhesus blood group-associated glycoprotein	0.67	0.96	0.000193	6p21.1-p11	164916	IncytePD:2048319
Hs.121521	ABL2	v-abl Abelson murine leukemia viral oncogene homolog 2	0.64	0.96	0.000889	1q24-q25	166612	IncytePD:1536149
Hs.112819		ESTs	1.41	0.96	0.000924	1	168969	IncytePD:2445101
Hs.277445	DGKZ	diacylglycerol kinase, zeta (104kD)	0.75	0.96	0.00194	11p11.2	159822	IncytePD:1875986
Hs.26915	SPTBN2	spectrin, beta, non-erythrocytic 2	0.63	0.95	0.000154	11q13	160846	IncytePD:1594108
Hs.12773	ACOX3	acyl-Coenzyme A oxidase 3, pristanoyl	0.74	0.95	0.000187	4p15.3	162487	IncytePD:3520054
Hs.79123	KIAA0084	KIAA0084 protein	0.7	0.95	0.000231	3p25.3-p25.1	159886	IncytePD:2697959

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.334841	SELENBP1	selenium binding protein 1	0.54	0.95	0.00043	1q21-q22	169315	IncytePD:2591494
Hs.2969	SKI	v-ski avian sarcoma viral oncogene homolog	1.65	0.95	0.000604	1q22-q24	164039	IncytePD:3283271
Hs.37129	SCNN1B	sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)	0.75	0.95	0.00083	16p12.2-p12.1	161191	IncytePD:1866654
Hs.25277	FLJ21065	hypothetical protein FLJ21065	1.25	0.94	7.57E-06	5	164202	IncytePD:2419078
Hs.83155	ALDH3B1	aldehyde dehydrogenase 3 family, member B1	0.58	0.94	8.76E-05	11q13	159838	IncytePD:2610218
Hs.24341	TAZ	transcriptional co-activator with PDZ-binding motif (TAZ)	1.27	0.94	9.26E-05	3q23-q24	164176	IncytePD:2345776
Hs.172458	IDS	iduronate 2-sulfatase (Hunter syndrome)	0.74	0.94	0.000322	Xq28	160243	IncytePD:549290
Hs.55279	SERPINB5	serine (or cysteine) proteinase inhibitor, member 5	0.62	0.94	0.00158	18q21.3	162215	IncytePD:460034
Hs.209587		ESTs, Weakly similar to I38022 hypothetical protein	1.58	0.94	0.00167	11	163251	IncytePD:1875433
Hs.118354	CAT56	CAT56 protein	0.69	0.93	4.71E-05	6	165027	IncytePD:3518549
Hs.182577	INPP5B	inositol polyphosphate-5-phosphatase, 75kD	0.74	0.93	0.000417	1p34	160074	IncytePD:1291948
Hs.81454	KHK	ketohexokinase (fructokinase)	0.65	0.93	0.000434	2p23.3-p23.2	159931	IncytePD:2516508
Hs.76688	CES1	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	0.18	0.93	0.000591	16q13-q22.1	164490	IncytePD:1813269
Hs.239499	KIAA0185	KIAA0185 protein	1.36	0.93	0.000722	10	168413	IncytePD:514653
Hs.151738	MMP9	matrix metalloproteinase 9 (gelatinase B, 92kD)	0.56	0.93	0.000722	20q11.2-q13.1	159912	IncytePD:1274074
Hs.186564		ESTs	1.38	0.93	0.000816	10	163409	IncytePD:1640094
Hs.198166	ATF2	activating transcription factor 2	0.68	0.93	0.00106	2q32	160057	IncytePD:2208152
Hs.149957	RPS6KA1	ribosomal protein S6 kinase, 90kD, polypeptide 1	0.75	0.93	0.00166	3	160006	IncytePD:1822236
Hs.36566	LIMK1	LIM domain kinase 1	0.6	0.92	4.11E-07	7q11.23	160082	IncytePD:3373632
Hs.50133		ESTs	1.52	0.92	6.67E-05	4	168567	IncytePD:1214652
Hs.14051		Homo sapiens mRNA; cDNA DKFZp434A2417	1.47	0.92	0.000246	10	168381	IncytePD:1431701
Hs.66731	HOXB13	homeo box B13	0.39	0.92	0.000572	17q21.2	159868	IncytePD:1861743
Hs.85087	LTBP4	latent transforming growth factor beta binding protein 4	0.68	0.92	0.000854	19q13.1-q13.2	159923	IncytePD:1956831

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.239706	GAB1	GRB2-associated binding protein 1	0.67	0.92	0.000933	4	162416	IncytePD:5066144
Hs.77554		cDNA FLJ14967 fis, similar to ZINC FINGER PROTEIN 84	4	0.92	0.0012	12	165454	IncytePD:1782052
Hs.14805	SLC21A11	solute carrier family 21 (organic anion transporter), member 11	1.49	0.92	0.00159	15q26	168293	IncytePD:408522
Hs.94498	LILRA2	leukocyte immunoglobulin-like receptor, subfamily A member 2	0.71	0.91	0.000459	19q13.4	161424	IncytePD:3336057
Hs.799	DTR	diphtheria toxin receptor (heparin-binding EGF-like growth factor)	0.61	0.91	0.000811	5q23	167412	IncytePD:1862257
Hs.28274		Homo sapiens cDNA: FLJ22049 fis, clone HEP09444	1.47	0.91	0.000856	8	163989	IncytePD:2155690
Hs.8358	FLJ20366	hypothetical protein FLJ20366	1.46	0.91	0.000952	8p22-q22.3	164145	IncytePD:3361529
Hs.293264		ESTs	1.38	0.91	0.00107	11	168371	IncytePD:829521
Hs.37953	FANCC	Fanconi anemia, complementation group C	0.62	0.91	0.00108	9q22.3	160036	IncytePD:3669589
Hs.250671	FLJ10140	hypothetical protein FLJ10140	1.47	0.91	0.00142	22q13	168397	IncytePD:642133
Hs.72964	MKRN3	makorin, ring finger protein, 3	0.63	0.91	0.00151	15q11-q13	164803	IncytePD:3181021
Hs.80683	MTRF1	mitochondrial translational release factor 1	1.24	0.91	0.00161	13q14.1-14.3	160533	IncytePD:1462246
Hs.79516	BASP1	brain abundant, membrane attached signal protein 1	0.62	0.9	2.60E-05	5p15.1-p14	159882	IncytePD:4008301
Hs.87539	ALDH3B2	aldehyde dehydrogenase 3 family, member B2	0.68	0.9	2.96E-05	11q13	166071	IncytePD:966447
Hs.38586	HSD3B1	hydroxy-delta-5-steroid dehydrogenase	0.64	0.9	0.000159	1p13.1	164787	IncytePD:182802
Hs.67688		ESTs	0.59	0.9	0.000311	6	162920	IncytePD:2789893
Hs.105584	RPS6KA4	ribosomal protein S6 kinase, 90kD, polypeptide 4	1.8	0.9	0.000409	11q11-q13	168189	IncytePD:2110163
Hs.24719	MAP-1	modulator of apoptosis 1	1.45	0.9	0.00108	14q32	168618	IncytePD:1967338
Hs.6232	KIAA0764	KIAA0764 gene product	1.36	0.9	0.00117	2pter-p25.1	163561	IncytePD:2043486
Hs.73792	CR2	complement component (3d/Epstein Barr virus) receptor 2	0.59	0.9	0.00181	1q32	160032	IncytePD:3055203
Hs.134342	LOC55915	TASP for testis-specific adriamycin sensitivity protein	1.44	0.9	0.00185	7q31.1-7q31.33	163421	IncytePD:1538396
Hs.33074		Homo sapiens, clone IMAGE:3606519, mRNA, partial cds	1.36	0.89	0.000275	8	168589	IncytePD:1431969
Hs.83795	IRF2	interferon regulatory factor 2	0.54	0.89	0.000848	4q34.1-q35.1	161188	IncytePD:2174666
Hs.81217	FZD2	frizzled (Drosophila) homolog 2	0.57	0.88	1.46E-06	17q21.1	160028	IncytePD:2214002
Hs.92357	GALK1	galactokinase 1	5.62	0.88	3.65E-05	17q24	169675	IncytePD:1215248
Hs.119273	KIAA0296	KIAA0296 gene product	0.53	0.88	6.98E-05	16p13.13-16p12.3	159951	IncytePD:3422646

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.194148	YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	0.64	0.88	9.16E-05	18p11.31-p11.21	159875	IncytePD:1887736
Hs.37054	EFNA3	ephrin-A3	0.69	0.88	0.000423	1q21-q22	161846	IncytePD:4178495
Hs.23643	MST4	serine/threonine protein kinase MASK	0.65	0.88	0.00108	X	163410	IncytePD:2793922
Hs.266959	HBG1	hemoglobin, gamma A	5.75	0.87	2.57E-05	11p15.5	168326	IncytePD:2156647
Hs.53478		Homo sapiens cDNA FLJ12366 fis, clone MAMMA1002411	1.34	0.87	5.78E-05	21	168383	IncytePD:1366043
Hs.283822	RHD	Rhesus blood group, D antigen	0.69	0.87	6.17E-05	1p36.2-p34.1	164821	IncytePD:1668024
Hs.118804	ENO3	enolase 3, (beta, muscle)	0.4	0.87	8.49E-05	17pter-p11	164468	IncytePD:1719955
Hs.772	GYS1	glycogen synthase 1 (muscle)	0.59	0.87	0.000112	19q13.3	160222	IncytePD:172916
Hs.77448	ALDH4A1	aldehyde dehydrogenase 4 family, member A1	0.66	0.87	0.00135	1p36	166147	IncytePD:831794
Hs.29640	RECK	reversion-inducing-cysteine-rich protein with kazal motifs	1.42	0.87	0.00172	9p13-p12	168569	IncytePD:2058483
Hs.93780		ESTs	1.12	0.87	0.00176		164377	IncytePD:2654539
Hs.11713	ELF5	E74-like factor 5 (ets domain transcription factor)	0.7	0.87	0.0018	11p13-p15	161000	IncytePD:2785892
Hs.97087	CD3Z	CD3Z antigen, zeta polypeptide (TiT3 complex)	0.6	0.85	0.000152	1q22-q23	160043	IncytePD:3227409
Hs.118795	FLJ10008	hypothetical protein FLJ10008	1.11	0.82	0.000317	14q22.1-q22.3	166653	IncytePD:2316425
Hs.17384		ESTs	1.16	0.82	0.000347	4	163225	IncytePD:2293931
Hs.1019	PTHR1	parathyroid hormone receptor 1	0.66	0.82	0.00102	3p22-p21.1	160109	IncytePD:1375235
Hs.77667	LY6E	lymphocyte antigen 6 complex, locus E	0.56	0.82	0.00145	8q24.3	162145	IncytePD:1472042
Hs.4988		Homo sapiens clone 24711 mRNA sequence	1.26	0.81	0.000274	2	160165	IncytePD:2061405
Hs.10669	DDEF1	development and differentiation enhancing factor 1	1.14	0.81	0.000896	8q24.1-q24.2	164026	IncytePD:2507108
Hs.5353	CASP10	caspase 10, apoptosis-related cysteine protease	1.01	0.81	0.00108	2q33-q34	164978	IncytePD:3984879
Hs.33102	TFAP2B	transcription factor AP-2 beta	0.58	0.81	0.00122	6p12	159845	IncytePD:2816550
Hs.144633	DKFZp434F232	hypothetical protein DKFZp434F2322	1.22	0.8	0.00132	17q24	163237	IncytePD:1473265
Hs.247423	ADD2	adducin 2 (beta)	0.55	0.79	0.000141	2p14-p13	162687	IncytePD:2112288
Hs.323712	KIAA0615	KIAA0615 gene product	1.3	0.79	0.00026	16q11.2-q12.2	163625	IncytePD:1217554
Hs.120360	PLA2G6	phospholipase A2, group VI (cytosolic, calcium-independent)	0.57	0.79	0.000778	22q13.1	160058	IncytePD:1849872

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.73800	SELP	selectin P (granule membrane protein 140kD, antigen CD62)	0.65	0.79	0.00156	1q22-q25	160049	IncytePD:3688202
Hs.65135	KIAA0913	KIAA0913 protein	1.16	0.78	0.00153	10	162465	IncytePD:2752015
		Unknown	0.99	0.77	0.000357		161881	IncytePD:2895226
Hs.274293		Homo sapiens mRNA; cDNA DKFZp761G1111	1.28	0.77	0.000542		165504	IncytePD:530360
Hs.153203	MDFI	MyoD family inhibitor	0.46	0.75	0.000138	6p21	163880	IncytePD:2645911
Hs.103393		ESTs	1.52	0.75	0.0014	16	163227	IncytePD:291636
Hs.153053	CD37	CD37 antigen	0.55	0.74	0.00022	19p13-q13.4	160033	IncytePD:3041162
Hs.98738	GRTH	gonadotropin-regulated testicular RNA helicase	1.06	0.72	0.000857	11q24	166657	IncytePD:2404557
Hs.180570	CYP4F12	cytochrome P450 isoform 4F12	0.55	0.72	0.0014	19p13.1	167601	IncytePD:1985566
Hs.50373		ESTs	5.25	0.7	2.91E-05	9	165500	IncytePD:372922
Hs.131705		ESTs	1.01	0.7	0.00128	8	165368	IncytePD:1921768
Hs.23672	LRP6	low density lipoprotein receptor-related protein 6	0.3	0.69	7.35E-05	12p11-p13	162040	IncytePD:4290851

Table 4. 30 Significant genes for predicting metastasis and their values necessary for computing multifactorial L value in the prediction model.

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.313	OPN	Osteopontin	1.07	3.29	0.00122	4	161923	IncytePD:4327691
Hs.69707	HCGII-7	HCGII-7 protein	1.07	2.85	0.000512	6	161462	IncytePD:1656490
Hs.177687	AKR1C4	aldo-keto reductase family 1, member C4	0.58	2.11	0.000939	10p15-p14	161753	IncytePD:5033671
		Unknown	0.82	1.74	0.0018		161371	IncytePD:3421817
Hs.276916	NR1D1	nuclear receptor subfamily 1, group D, member 1	0.74	1.71	0.00181	17q11.2	166707	IncytePD:1904760
Hs.211569	GPRK5	G protein-coupled receptor kinase 5	0.99	1.69	0.00147	10q24-qter	161133	IncytePD:1418741
Hs.75573	CENPE	centromere protein E (312kD)	0.82	1.65	1.00E-06	4q24-q25	160128	IncytePD:3081067
Hs.283664	ASPH	aspartate beta-hydroxylase	0.7	1.56	0.000576	8q12.1	160084	IncytePD:3693273
Hs.296371	RAB28	RAB28, member RAS oncogene family	1.07	1.5	0.000833	4p16.1	160699	IncytePD:1457948
Hs.274313	IGFBP6	insulin-like growth factor binding protein 6	0.87	1.21	0.00192	12q13	160319	IncytePD:1968126
Hs.34526	TYMSTR	G protein-coupled receptor	0.89	1.18	0.00101	3p21	161635	IncytePD:2610374
Hs.222	ITGA9	integrin, alpha 9	0.69	1.16	3.74E-06	3p21.3	160135	IncytePD:2487318
Hs.63984	CDH13	cadherin 13, H-cadherin	0.72	1.13	0.000103	16q24.2-q24.3	160122	IncytePD:1404153
Hs.75596	IL2RB	interleukin 2 receptor, beta	0.69	1	0.000688	22q13.1	159942	IncytePD:3936210
Hs.55279	SERPINB5	serine (or cysteine) proteinase inhibitor, member 5	0.62	0.94	0.00158	18q21.3	162215	IncytePD:460034
Hs.118354	CAT56	CAT56 protein	0.69	0.93	4.71E-05	6	165027	IncytePD:3518549
Hs.182577	INPP5B	inositol polyphosphate-5-phosphatase, 75kD	0.74	0.93	0.000417	1p34	160074	IncytePD:1291948
Hs.81454	KHK	ketohexokinase (fructokinase)	0.65	0.93	0.000434	2p23.3-p23.2	159931	IncytePD:2516508
Hs.76688	CES1	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	0.18	0.93	0.000591	16q13-q22.1	164490	IncytePD:1813269

UG cluster	Name	Description	PN	PT	p value	map	Unique id	Clone
Hs.151738	MMP9	matrix metalloproteinase 9 (gelatinase B, 92kD)	0.56	0.93	0.000722	20q11.2-q13.1	159912	IncytePD:1274074
Hs.94498	LILRA2	leukocyte immunoglobulin-like receptor, subfamily A member 2	0.71	0.91	0.000459	19q13.4	161424	IncytePD:3336057
Hs.83795	IRF2	interferon regulatory factor 2	0.54	0.89	0.000848	4q34.1-q35.1	161188	IncytePD:2174666
Hs.81217	FZD2	frizzled (Drosophila) homolog 2	0.57	0.88	1.46E-06	17q21.1	160028	IncytePD:2214002
Hs.194148	YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	0.64	0.88	9.16E-05	18p11.31-p11.21	159875	IncytePD:1887736
Hs.23643	MST4	serine/threonine protein kinase MASK	0.65	0.88	0.00108	X	163410	IncytePD:2793922
Hs.118804	ENO3	enolase 3, (beta, muscle)	0.4	0.87	8.49E-05	17pter-p11	164468	IncytePD:1719955
Hs.153203	MDFI	MyoD family inhibitor	0.46	0.75	0.000138	6p21	163880	IncytePD:2645911
Hs.153053	CD37	CD37 antigen	0.55	0.74	0.00022	19p13-q13.4	160033	IncytePD:3041162
Hs.180570	CYP4F12	cytochrome P450 isoform 4F12	0.55	0.72	0.0014	19p13.1	167601	IncytePD:1985566
Hs.23672	LRP6	low density lipoprotein receptor-related protein 6	0.3	0.69	7.35E-05	12p11-p13	162040	IncytePD:4290851

B. Example 2: Predicting a predisposition for Hepatocellular Carcinoma

1. Material and methods

a) Patients and tissue samples

[0243] Surgical specimens were collected with prior informed consent and with the protocols and the approval by the Institution Review Board of University of Minnesota. Liver samples were obtained from 59 end-stage chronic liver disease patients who received liver transplantation between 1995-2001. Disease-free liver samples from 8 liver donors were used as control. The collection of these samples was mainly managed through the Liver Tissue Procurement and Distribution System (LTPADS) at University of Minnesota, USA. Tumor and matched non-tumor liver samples from 64 patients were obtained through either the LTPADS program or Liver Cancer Institute at Fudan University, China. Frozen samples once received was stored immediately at –80°C in a tissue repository database.

b) cDNA microarray

[0244] Total RNA was extracted from frozen tissues by using Trizol reagent (Invitrogen, Gaithersburg, MD) according to the manufacturer's protocol. The quality of extracted RNA was determined by spectrophotometry and by the appearance of characteristic 28S and 18S rRNA fragment on a 1% agarose gel. Each RNA sample divided into several tubes same amount and stored –80°C. For the common reference of cDNA microarray, total RNA samples from 8 normal liver were combined together, and were aliquoted into each tubs.

[0245] cDNA microarrays were purchased from NCI microarray facility, Advanced Technology Center, NCI, NIH (Gaithersburg, MD). These human UniGem v2.0 array contained 9180 cDNA clones that map into 8281 unique UniGene clusters (base on Hs Unigene Build #131 released on Feb. 28, 2001) and 122 Incyte EST clones (Incyte Genomics, Palo Alto, CA). The hybridization was performed according to an optimized protocol established by the NCI (Wu et al, *Oncogene* **20**:3674-3682, 2001; Ye et al, *Nature Med.* *9*:416-423, 2003). Fluorescent images of hybridized microarrays were obtained by using GenePix 4000 scanner and GenePix Pro software (Axon Instruments, Foster City, CA). Detailed information as being collected according to the proposed Minimum Information About a Microarray Experiment Standards (Brazma A et al., Nat Genet 2001) will be made available through the NCBI's Gene Expression Ominibus public database.

c) Statistical analysis

[0246] A hierarchical clustering analysis was preformed using a relative gene expression ratio (Cy5/Cy3) to examine the relatedness among expression patterns of several gene lists and those in two risk groups. Cluster analysis was performed using Cluster software and visualized using Tree View software (Eisen et al., supra). Hierarchical clustering was performed following median centering normalization.

[0247] Analyses were performed using BRB ArrayTools developed by Dr. Richard Simon and Amy Peng of the Biometrics Research Branch at National Cancer Institute. The data from each array were scaled in order to normalize data for inter-array comparisons. The class comparison tool was used for comparing two pre-defined risk groups. The F-test was a generalization of the two-sample t-test for comparing values among groups. The class comparison tool computed an F-test separately for each gene using the normalized log-ratios for cDNA. Several other important statistics were also computed. The tool performed random permutations of the group. Based on these random permutations, the tool computed the permutation p value associated with each gene in the list.

[0248] Classification of samples into one of two pre-determined classes based on gene expression data was performed using several algorithms including compound covariate predictor, K-nearest neibougher predictor, or support vector machine predictor. The predictor was built in two steps. First, a standard two-sample *t*-test was performed to identify genes with significant differences (at level 0.001) in log-expression ratios between the two classes. Second, the log-expression ratios of differentially expressed genes were combined into a single compound covariate for each sample; the compound covariate was used as the basis for class prediction. The compound covariate for sample *i* was defined as

$$c_i = \sum_j t_j x_{ij},$$

where t_j was the *t*-statistic for the two group comparison of classes with respect to gene *j*, x_{ij} was the log-ratio measured in specimen *i* for gene *j* and the sum is over all differentially expressed genes.

[0249] We predicted the classification of a new sample by computing the following linear combination:

$$L = \Sigma_i\ t_i *(x_i - m_i)$$

where t_i was t-value for gene *i*, x_i was log-ratio of gene *i* in the new sample to be classified, and m_i was midpoint between the two classes for gene *i*. The index *i* run over all the genes that are significant in the original analysis. When L was positive, then the new sample should be classified to be of the first phenotype label whereas L was negative, then the new sample should be classified to be of the second phenotype label.

d) EpCAM expression and its in vitro inhibition

[0250] The expression of EpCAM was assessed by semi-quantitative PCR. Total RNA was reversed-transcribed to produce single-stranded cDNA using random primers (Promega) with Superscript II reverse transcriptase (Invitrogen) according to manufacturer's protocol. PCR amplification was performed with QuantumRNA 18S Internal Standards (Ambion) by using HotStarTaq DNA polymerase (Qiagen) according to manufacturer's protocol. The primer sequences are as follow: forward, 5'-TGC CGC AGC TCA GGA AGA ATG TGT-3' (SEQ ID NO:6); reverse, 5'-CAT CAT TCT GAG TTT TTT GAG AAG-3' (SEQ ID NO:7).

[0251] siRNA was used to inhibit EpCAM expression. siRNA were synthesized by Qiagen. The sense and antisence strands of EpCAM are: sense, 5'-GUU UGC GGA CUG CAC UUC AdTdT-3' (SEQ ID NO:8); antisense, 5'-UGA AGU GCA GUC CGC AAA CdTdT-3' (SEQ ID NO:9). Non-silencing RNA was purchased from Qiagen and used as control siRNA. The sequences of control siRNA were: sense, 5'-UUC UCC GAA CGU GUC ACG UdTdT-3' (SEQ ID NO:10); antisense, 5'-ACG UGA CAC GUU CGG AGA AdTdT-3' (SEQ ID NO:11). Transfection of siRNAs was carried out using TransIT-TKO transfection reagent (Mirus) according to the manufacturer's protocol and 200 nM siRNA duplex per experiment. Cell growth was determined by using Cell Counting Kit-8 (Dojindo Molecular Tech.) as described by the manufacturer. The experiments were performed in triplicate.

2. Results

[0252] Gene expression profiles of liver samples from 59 chronic liver disease (CLD) patients and of 14 HCC samples were compared to that of a pool of 8 disease-free normal liver samples by microarray containing 9128 human cDNA clones (Ye et al., *Supra*). The CLD samples included 7 hepatitis B (HBV), 11 hepatitis C (HCV), 3 hemochromatosis (HHC), 5 Wilson's Disease (WD), 10 alcoholic liver disease (ALD), 16 primary biliary cirrhosis (PBC) and 7 autoimmune hepatitis (AIH). A supervised univariate F-test algorithm with 2000 random permutations of the class labels was used to search for genes that can discriminate these 7 CLD groups. This analysis yielded a total of 489 significant genes ($p<0.0005$). Hierarchical clustering analysis (as described by Eisen et al., supra) of the 489 genes revealed that these 7 liver disease groups were separated into two major branches, one consisting mostly of HBV, HCV, HHC, and WD samples and other containing mainly PBC, ALD, and AIH samples. These results indicate that HBV, HCV, HHC, and WD are more closely related each other than they are as a group to PBC, ALD, or AIH. The segregation of these samples by a molecular signature specifically reflecting their etiologies was correlated coincidentally with their risk to develop HCC, with an exception of WD samples (data not shown). To further determine the degree of difference among these groups, a t-test was performed based compound covariate predictor analysis among these 7 groups with "leave-one-out" cross-validation and 2000 random permutation tests. A total of 21 simulations were performed, which yielded 500 composite genes. The result of the hierarchical clustering of these genes is consistent with that of F-test (data not shown). Consistently, PBC, ALD, or AIH was more significantly different from HBV, HCV, HHC, or WD, while the differences among the etiologies were less significant (data not shown).. It appears that the WD samples, at least for this set, belong to the high-risk group. The interpretations from above results are that the molecular signature is dominated by the genes segregating the high risk group from the low risk group for their ability to develop HCC while genes reflecting their individual etiologies were minuscule.

[0253] The genes that were commonly disregulated in HBV/HCV/HHC/WD samples but not in ALD/PBC/AIH were hypothesized to be more closely related to the molecular signature of HCC. To search globally for such a gene set, the k-nearest neighbors (K=3) (3NN) and support vector machine (SVM) algorithms were applied with a "leave-one-out" cross-validation test and 2000 random permutations of class labeling test to the high risk (HBV/HCV/HHC/WD) and low risk (ALD/PBC/AIH) groups at a P value <0.001, a computation strategy similar to our recent study (Ye et al., *supra*). This analysis yielded a composite classifier containing 556 significant genes, which separated these two groups very well. It provided a significant class prediction among these groups with an overall accuracy of 78% by 3NN and 86% by SVM, respectively, and the cross-validated misclassification rates were significantly lower than expected by chance (p<0.0005) (data not shown). However, random grouping of these samples yielded statistically insignificant classification (data not shown).

[0254] It was noted that many genes in the 556-gene set can be found in the 14 HCC samples analyzed (data not shown). To identify genes that were commonly disregulated in the high-risk group and in HCC, the 14 HCC samples were pooled together with the high-risk group and then compared with the low risk group using 3NN algorithm at a *P* value <0.001, with 2000 random permutations. This analysis yielded 416 genes, in which 273 genes were found in the 556-gene set (49% overlapping). These results indicate that about half of the signature genes that can discriminate between the high risk and the low risk groups are present in HCC samples. To determine if the 273-gene set (Table 5) was a common signature for tumors, we applied this set to two independent HCC gene expression profiles using the 3NN and SVM predictors. One set included 24 HCC samples derived from a comparison with the same normal liver control used above and the other set including 50 HCC samples that were compared to its matched non-cancerous liver tissues (Ye et al., *supra*). The 273-gene signature provided an increased fitness by SVM in their classification with an overall accuracy of 92% for the 24 HCC samples and 94% for the 50 HCC samples (data not shown), which was improved in overall performance as compared to the 556-gene set. Consistently, the non-overlapping 283-gene set did not provide any satisfactory performance. Because most of the HCC-associated genes in the non-overlapping gene set were eliminated, most of the 283 genes may belong to the signatures separating the etiologies. Moreover, the 383 overlapping genes selected from a comparison of HBV/HCV/HHC/WD and ALD/PBC/AIH/HCC did not yield a meaningful classification of the two independent HCC sets with an overall predictive rate below 50% (a random event). The 273 genes were examined in multiple liver samples taken from two HBV patients and from different parts of the liver that were spread at least in a 5 cm diameter region. The profiles of these 273 genes in different parts of the livers from these two patients were almost identical (data not shown). Furthermore, top 25 genes with the lowest parametric p-values (p<0.000001) were selected from the 273-gene set. This set gave rise to a comparable result as the 273-gene set (data not shown). Taken together, these results indicate that the 273-gene set contains most of the HCC-associated genes relevant to HCC development and that these genes are widely spread in the parenchyma of the affected livers rather than are retained locally.

[0255] To examine if the 273-gene set is a common signature in other human tumors, the gene parameters in this signature were applied using SVM to 98 HCCs, 53 lung cancers, 89 gastric adenocarcinoma, 37 soft tissue tumors, 39 breast tumors and 23 difuse large B-cell lymphoma (DLBCL) from several publicly available microarray datasets (Alizadeh et al., *supra*; Perou et al., *supra*; Garber et al., *Proc. Natl Acad. Sci. U.S.A.* 98:13784-13789, 2001). While the 273-gene set consistently performed well with additional 98 HCC samples (80% of the samples fit the signature), 97% of breast cancers (39 cases) and 78% of DLBCL cases shared similar signatures. In contrast, most of the tumor samples from lung, soft tissues, and stomach showed a very poor fit to this signature (between 6 and 30% of the cases) (data not shown). As a control, the 283-gene set (non-HCC-related genes) did not provide a satisfying prediction to these samples. Thus, the HCC-associated genes in the classifier appear to be commonly disregulated in breast cancer and DLBCL, but not in lung adenocarcinoma, soft tissue tumors, and gastric adenocarcinoma.

[0256] Above studies suggested that genes responsible for the genesis of HCC may be present in the 273 gene set. For example, the gene whose expression is significantly elevated in the high-risk group but not in the low-risk group may act as an oncogene to promote cell growth. To test this "proof-of-principle" hypothesis, a lead gene at the top of the 273 genelist was selected. This gene was identified as EpCAM or tumor-associated calcium signal transducer 1 (TACSTD1, Hs.692), with an average of a 3.6-fold increased expression in the high risk group but only a 1.7 fold in the low risk group (Fig 6a) as well as in HCC (data not shown). Elevated expressions of EpCAM in the high-risk CLD samples were verified by the quantitative RT-PCR analysis (Fig 6b). The expression of EpCAM in various HCC cell lines was examined by Western blot analysis. EpCAM is highly expressed in Hep3B cells but the expression level is relatively low in Huh1 and Huh4 cells (Fig 6c), generally correlating with their growth rates (Fig 6d). Furthermore, inhibition of EpCAM expression by two different siRNA oligos specific to EpCAM resulted in a significant growth inhibition of Hep3B cells (Fig 6f). In contrast, a control siRNA oligo has no such effect (Fig 6e and data not shown). These results indicate that EpCAM may provide oncogenic property by promoting neoplastic cell proliferation.

[0257] The 273 significant genes, their gene symbols, their map positions, and their UG Cluster identifiers are presented in Table 5.

Table 5. 273 significant genes for predicting the potential for developing HCC in a patient with a chronic liver disease and their values necessary for computing multifactorial L value in the prediction model.

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
1	-7.28	p < 0.000001	0.603	0.903	160198	cofilin 2 (muscle)	Hs.180141	CFL2	14q
2	-6.53	p < 0.000001	0.985	1.607	168023	Fc fragment of IgG, high affinity Ia, receptor for (CD64)	Hs.77424	FCGR1A	1q21.2-q21.3
3	-6.46	p < 0.000001	0.643	1.175	162315	calcium channel, voltage-dependent, beta 3 subunit	Hs.250712	CACNB3	12q13
4	-6.18	p < 0.000001	0.688	1.112	160302	myosin IB	Hs.121576	MYO1B	2q12-q34
5	-6.16	p < 0.000001	0.473	1.161	169417	ceruloplasmin (ferroxidase)	Hs.296634	CP	3q23-q25
6	-6.1	p < 0.000001	0.876	1.18	161756	albumin	Hs.184411	ALB	4q11-q13
7	-6.04	p < 0.000001	0.719	1.224	162290	UDP-N-acteylglucosamine pyrophosphorylase 1	Hs.21293	UAP1	1q23.1
8	-6.01	p < 0.000001	0.534	1.141	162538	Unknown [Homo sapiens], mRNA sequence	Hs.367982		16
9	-5.94	p < 0.000001	0.491	0.714	168634	chromosome 20 open reading frame 3	Hs.22391	C20orf3	20p11.22-p11.21
10	-5.93	p < 0.000001	0.756	1.276	164136	acyl-Coenzyme A dehydrogenase, long chain	Hs.1209	ACADL	2q34-q35
11	-5.9	p < 0.000001	0.864	1.181	163874	KIAA0092 gene product	Hs.151791	KIAA0092	11q21
12	-5.88	p < 0.000001	0.728	0.925	163096	CGI-26 protein	Hs.24332	CGI-26	12p12.3

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
13	-5.73	p < 0.000001	0.616	1.133	160233	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 3	Hs.38018	DYRK3	1q32
14	-5.67	p < 0.000001	0.786	1.071	160436	Similar to hypothetical protein PRO2831 [Homo sapiens], mRNA sequence	Hs.406646		15
15	-5.65	p < 0.000001	0.761	1.382	160795	hepatic leukemia factor	Hs.433707	HLF	17q22
16	-5.61	p < 0.000001	0.314	0.798	161944	complement component 9	Hs.1290	C9	5p14-p12
17	-5.6	p < 0.000001	0.506	0.703	167718	ATP-binding cassette, sub-family A (ABC1), member 1	Hs.211562	ABCA1	9q31.1
18	-5.58	p < 0.000001	0.65	0.912	168437	KIAA0843 protein	Hs.26777	KIAA0843	5q32
19	-5.57	p < 0.000001	0.843	1.087	162884	intracellular membrane-associated calcium-independent phospholipase A2 gamma	Hs.44198	IPLA2(GAMMA)	7q31
20	-5.48	p < 0.000001	0.657	1.065	166910	SIPL protein	Hs.64322	SIPL	2p25.3
21	-5.46	1.00E-06	0.544	1.003	166192	ESTs, Highly similar to MT1B_HUMAN METALLOTHIONEIN-IB (MT-1B) [H.sapiens]	Hs.36102		16
22	-5.46	1.00E-06	0.46	0.832	164779	N-acetyltransferase 2 (arylamine N-acetyltransferase)	Hs.2	NAT2	8p22
23	-5.44	1.00E-06	0.707	1.191	166252	CD5 antigen-like	Hs.52002	CD5L	1q21-q23

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
						(scavenger receptor cysteine rich family)			
24	-5.44	1.00E-06	0.861	1.512	162878	ESTs, Highly similar to alpha 1 type XI collagen, isoform B preproprotein; collagen XI, alpha-1 polypeptide [Homo sapiens] [H.sapiens]	Hs.7967		1
25	-5.42	1.00E-06	0.767	1.181	164656	N-chimaerin (AA 1-299) [Homo sapiens], mRNA sequence	Hs.385460		2
26	-5.42	2.00E-06	0.803	1.296	161780	Incyte EST	3441835 (IncytePD)		
27	-5.38	1.00E-06	0.352	0.745	160174	complement component 6	Hs.1282	C6	5p13
28	-5.35	2.00E-06	0.464	0.875	160280	carboxypeptidase B2 (plasma, carboxypeptidase U)	Hs.75572	CPB2	13q14.11
29	-5.34	2.00E-06	0.779	0.978	163144	KIAA1724 protein	Hs.127243	KIAA1724	2p23.3
30	-5.33	2.00E-06	0.694	1.361	169477	mannose receptor, C type 1	Hs.75182	MRC1	10p13
31	-5.26	2.00E-06	0.669	0.896	162659	RAB6A, member RAS oncogene family	Hs.5636	RAB6A	11q13.3
32	-5.25	2.00E-06	0.768	1.052	161138	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	Hs.297681	SERPINA1	14q32.1

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
33	-5.25	3.00E-06	0.685	1.043	169635	ESTs, Weakly similar to ubiquitously transcribed tetratricopeptide repeat gene, Y chromosome; Ubiquitously transcribed TPR gene on Y chromosome [Homo sapiens] [H.sapiens]	Hs.87980		2
34	-5.21	3.00E-06	0.598	0.815	162745	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1	Hs.91139	SLC1A1	9p24
35	-5.2	3.00E-06	0.371	0.725	160366	ubiquitin specific protease 10	Hs.78829	USP10	16q24.1
36	-5.16	3.00E-06	0.515	0.932	166426	protein S (alpha)	Hs.64016	PROS1	3p11-q11.2
38	-5.14	4.00E-06	0.627	1.044	162301	interleukin 1 receptor accessory protein	Hs.173880	IL1RAP	3q28
39	-5.11	4.00E-06	0.534	0.919	167159	steroid-5-alpha-reductase, alpha polypeptide 2 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 2)	Hs.1989	SRD5A2	2p23
40	-5.04	5.00E-06	0.474	0.9	167129	metallothionein 1L	Hs.380778	MT1L	16q13
41	-5.02	5.00E-06	0.87	2.237	163633	leptin receptor	Hs.226627	LEPR	1p31
42	-5.02	5.00E-06	0.506	1.137	162311	serine (or cysteine) proteinase inhibitor, clade C (antithrombin), member 1	Hs.75599	SERPINC1	1q23-q25.1

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
43	-5.01	6.00E-06	0.622	1.035	166915	hypothetical protein FLJ12666	Hs.23767	FLJ12666	1p34.2
44	-5	6.00E-06	0.741	1.14	163572	hypothetical protein DKFZp564D0462	Hs.44197	DKFZP564D046	6q23.1-q24.3
45	-5	6.00E-06	0.842	1.141	163676	inositol(myo)-1(or 4)-monophosphatase 1	Hs.171776	IMPA1	8q21.13-q21.3
46	-5	6.00E-06	0.903	1.145	163549	ESTs, Weakly similar to ARF protein [Homo sapiens] [H.sapiens]	Hs.422650		17
47	-4.99	6.00E-06	0.357	0.608	168690	corticotropin releasing hormone binding protein	Hs.115617	CRHBP	5q11.2-q13.3
48	-4.99	6.00E-06	0.52	0.846	169399	pregnancy-zone protein	Hs.74094	PZP	12p13-p12.2
49	-4.98	6.00E-06	0.681	0.994	162636	signal recognition particle 54kDa	Hs.49346	SRP54	14q13.1
50	-4.98	6.00E-06	0.633	0.933	166021	inositol polyphosphate-5-phosphatase, 145kDa	Hs.155939	INPP5D	2q36-q37
51	-4.93	7.00E-06	0.972	1.427	159896	neural precursor cell expressed, developmentally down-regulated 4	Hs.1565	NEDD4	15q
52	-4.92	8.00E-06	0.73	1.087	163778	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1	Hs.20894	NDST1	5q32-q33.1
53	-4.9	8.00E-06	0.705	1.067	159807	kidney ankyrin repeat-containing protein	Hs.77546	KANK	9p24.3

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
54	-4.9	8.00E-06	0.307	0.676	167252	hydroxyprostaglandin dehydrogenase 15-(NAD)	Hs.77348	HPGD	4q34-q35
55	-4.88	9.00E-06	0.724	1.417	163254	lipase A, lysosomal acid, cholesterol esterase (Wolman disease)	Hs.85226	LIPA	10q23.2-q23.3
56	-4.87	1.00E-05	0.576	0.923	162307	protein-L-isoaspartate (D-aspartate) O-methyltransferase	Hs.79137	PCMT1	6q24-q25
57	-4.87	9.00E-06	0.64	1.076	164602	complement component 1, s subcomponent	Hs.169756	C1S	12p13
58	-4.83	1.10E-05	1.057	1.872	164576	forkhead box O1A (rhabdomyosarcoma)	Hs.170133	FOXO1A	13q14.1
59	-4.8	1.20E-05	0.78	1.259	165739	hypothetical gene CG018	Hs.22174	CG018	13q12-q13
60	-4.8	1.20E-05	0.719	1.091	167087	solute carrier family 31 (copper transporters), member 2	Hs.24030	SLC31A2	9q31-q32
61	-4.79	1.20E-05	0.716	0.987	165277	phosphorylase, glycogen; liver (Hers disease, glycogen storage disease type VI)	Hs.771	PYGL	14q21-q22
62	-4.7	1.70E-05	0.766	1.43	161801	solute carrier family 10 (sodium/bile acid cotransporter family), member 1	Hs.952	SLC10A1	14q24.1
63	-4.7	1.80E-05	0.355	0.917	162617	FK506 binding protein 5	Hs.7557	FKBP5	6p21.3-21.2
64	-4.68	1.80E-05	0.918	1.294	163597	hypothetical protein	Hs.8358	FLJ20366	8q23.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
						FLJ20366			
65	-4.67	1.90E-05	0.598	0.848	160741	aldehyde dehydrogenase 8 family, member A1	Hs.18443	ALDH8A1	6q23.2
66	-4.67	1.90E-05	0.392	0.742	167158	complement component 5	Hs.1281	C5	9q32-q34
67	-4.65	2.00E-05	1.003	1.661	165565	phosphatidylinositol (4,5) bisphosphate 5-phosphatase homolog; phosphatidylinositol polyphosphate 5-phosphatase type IV	Hs.25156	PPI5PIV	9q34.3
68	-4.65	2.00E-05	0.996	1.169	160476	likely ortholog of mouse deleted in polyposis 1	Hs.178112	DP1	5q22-q23
69	-4.65	2.50E-05	0.93	1.204	161778	protein phosphatase 1D magnesium-dependent, delta isoform	Hs.100980	PPM1D	17q23.2
70	-4.62	2.20E-05	0.875	1.013	164997	N-acetylgalactosaminidase, alpha-	Hs.75372	NAGA	22q13-qter
71	-4.62	2.30E-05	1.04	1.351	160731	histone deacetylase 6	Hs.6764	HDAC6	Xp11.23
72	-4.62	2.30E-05	0.98	1.326	168995	ring finger protein 13	Hs.6900	RNF13	3q25.1
73	-4.6	2.40E-05	0.536	0.805	163500	coagulation factor XI (plasma thromboplastin antecedent)	Hs.1430	F11	4q35
74	-4.59	2.50E-05	0.359	0.544	159810	C-type lectin BIMLEC precursor	Hs.2441	BIMLEC	2q24.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
75	-4.57	2.60E-05	0.912	1.66	168655	complement component 1, q subcomponent, beta polypeptide	Hs.8986	C1QB	1p36.3-p34.1
76	-4.57	2.70E-05	0.529	1.031	166497	histidine ammonia-lyase	Hs.276590	HAL	12q22-q24.1
77	-4.57	3.60E-05	0.421	0.88	161748	acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)	Hs.37	ACAT1	11q22.3-q23.1
78	-4.56	2.70E-05	0.636	1.205	164394	CD163 antigen	Hs.74076	CD163	12p13.3
79	-4.54	2.90E-05	0.926	1.178	160011	general transcription factor IIA, 2, 12kDa	Hs.76362	GTF2A2	15q21.3
80	-4.54	3.10E-05	0.634	0.922	161895	nuclear receptor subfamily 1, group I, member 2	Hs.118138	NR1I2	3q12-q13.3
81	-4.54	3.00E-05	0.907	1.181	167754	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 926491, mRNA sequence	Hs.98401		19
82	-4.54	4.10E-05	0.988	1.3	161838	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 1, 6kDa	Hs.84549	NDUFC1	4q28.2-q31.1
83	-4.47	3.70E-05	1.124	1.642	161856	glutathione-S-transferase like; glutathione transferase omega	Hs.11465	GSTTLp28	10q24.33
84	-4.47	3.80E-05	0.893	1.216	163456	phytanoyl-CoA hydroxylase (Refsum disease)	Hs.172887	PHYH	10pter-p11.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
85	-4.46	3.90E-05	0.51	0.865	168256	B-factor, properdin	Hs.69771	BF	6p21.3
86	-4.43	4.30E-05	0.611	1.011	162472	angiogenin, ribonuclease, RNase A family, 5	Hs.332764	ANG	14q11.1-q11.2
87	-4.41	4.80E-05	0.593	0.906	167629	N-acetyltransferase 1 (arylamine N-acetyltransferase)	Hs.155956	NAT1	8p23.1-p21.3
88	-4.39	5.90E-05	0.884	1.231	162036	Dombrock blood group	Hs.13776	DO	12q13.2-q13.3
90	-4.39	5.00E-05	0.448	0.831	159972	pre-B-cell colony-enhancing factor	Hs.239138	PBEF	7q22.1
91	-4.38	5.10E-05	0.892	1.14	160759	glucuronidase, beta	Hs.183868	GUSB	7q21.11
92	-4.37	5.20E-05	0.797	1.284	162192	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	Hs.79158	ACADM	1p31
93	-4.37	5.40E-05	0.811	1.062	161636	Homo sapiens clone 24405 mRNA sequence	Hs.23729		1
94	-4.34	5.80E-05	0.746	1.211	168452	methylenetetrahydrofolate dehydrogenase (NADP+ dependent), methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase	Hs.172665	MTHFD1	14q24
95	-4.33	6.10E-05	0.541	0.906	165666	ribonuclease, RNase A family, 4	Hs.283749	RNASE4	14q11.1
96	-4.33	6.20E-05	0.482	0.939	167394	butyrylcholinesterase	Hs.1327	BCHE	3q26.1-

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
									q26.2
97	-4.3	6.80E-05	0.62	0.767	167501	propionyl Coenzyme A carboxylase, alpha polypeptide	Hs.80741	PCCA	13q32
98	-4.3	6.80E-05	0.809	2.181	165974	insulin-like growth factor binding protein 1	Hs.102122	IGFBP1	7p13-p12
99	-4.29	7.00E-05	0.622	0.933	161234	plakophilin 2	Hs.25051	PKP2	12p11
100	-4.29	7.00E-05	0.852	1.098	166532	phosphatidylcholine transfer protein	Hs.285218	PCTP	17q21-q24
101	-4.28	7.40E-05	0.567	0.815	167750	adenosine kinase	Hs.432422	ADK	10cen-q24
102	-4.27	7.80E-05	0.479	0.766	165890	fibrinogen, B beta polypeptide	Hs.7645	FGB	4q28
103	-4.26	7.70E-05	0.406	0.89	161362	tryptophan 2,3-dioxygenase	Hs.183671	TDO2	4q31-q32
104	-4.25	8.00E-05	0.739	1.044	159764	annexin A7	Hs.386741	ANXA7	10q21.1-q21.2
105	-4.25	8.10E-05	0.642	0.88	164249	aminocarboxymuconate semialdehyde decarboxylase	Hs.114088	ACMSD	2q21.2
106	-4.24	8.30E-05	0.91	1.142	162711	mitofusin 2	Hs.3363	MFN2	1p36.21
107	-4.24	8.30E-05	0.784	1.391	160370	serum/glucocorticoid regulated kinase	Hs.296323	SGK	6q23
108	-4.24	8.40E-05	0.483	0.867	161146	3-hydroxysteroid epimerase	Hs.11958	RODH	12q13

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
109	-4.23	9.10E-05	0.476	0.846	161986	tumor rejection antigen (gp96) 1	Hs.82689	TRA1	12q24.2-q24.3
110	-4.23	8.60E-05	0.807	1.049	165670	toll-like receptor 2	Hs.63668	TLR2	4q32
111	-4.22	8.80E-05	0.577	0.78	166820	KIAA0212 gene product	Hs.154332	KIAA0212	3p26.1
112	-4.21	9.10E-05	0.604	0.838	164495	Homo sapiens, clone IMAGE:3833472, mRNA, mRNA sequence	Hs.234898		12
113	-4.21	9.10E-05	0.407	0.592	163893	fibrinogen-like 1	Hs.107	FGL1	8p22-p21.3
114	-4.2	9.30E-05	0.651	1.058	167287	cytochrome b-5	Hs.83834	CYB5	18q23
115	-4.2	9.40E-05	0.597	1.015	162446	electron-transferring-flavoprotein dehydrogenase	Hs.323468	ETFDH	4q32-q35
116	-4.19	9.90E-05	0.507	1.102	169375	cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 9	Hs.167529	CYP2C9	10q24
117	-4.18	0.000103	0.523	0.963	160720	sorbitol dehydrogenase	Hs.878	SORD	15q15.3
118	-4.17	0.000107	0.992	1.266	162067	splicing factor 3b, subunit 1, 155kDa	Hs.334826	SF3B1	2q33.1
119	-4.15	0.000115	0.639	0.936	164393	Homo sapiens mRNA; cDNA DKFZp762O1615 (from clone DKFZp762O1615), mRNA sequence	Hs.284252		5

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
120	-4.15	0.000114	0.794	1.029	162329	estrogen receptor binding site associated, antigen, 9	Hs.9222	EBAG9	8q23
121	-4.14	0.000116	0.59	1.176	164863	solute carrier family 2 (facilitated glucose transporter), member 2	Hs.167584	SLC2A2	3q26.1-q26.2
122	-4.14	0.000117	0.767	1.029	163052	fused toes homolog (mouse)	Hs.288929	FTS	16q12.1
123	-4.12	0.000124	0.712	0.997	160399	cullin 3	Hs.78946	CUL3	2q36.3
124	-4.12	0.000124	0.649	0.837	165894	protein kinase, cAMP-dependent, regulatory, type II, beta	Hs.77439	PRKAR2B	7q22-q31.1
125	-4.11	0.000126	0.941	1.258	162938	PTD013 protein	Hs.22679	PTD013	6q13-q22.33
126	-4.09	0.000137	0.622	0.958	160328	pre-alpha (globulin) inhibitor, H3 polypeptide	Hs.76716	ITIH3	3p21.2-p21.1
127	-4.08	0.000142	0.718	1.057	165794	epoxide hydrolase 2, cytoplasmic	Hs.113	EPHX2	8p21-p12
128	-4.07	0.000149	0.405	0.709	162561	RNA helicase-related protein [Homo sapiens], mRNA sequence	Hs.381097		16
129	-4.06	0.000149	0.447	0.743	168811	acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)	Hs.37	ACAT1	11q22.3-q23.1
130	-4.06	0.000152	0.949	1.293	169563	zinc finger protein 103 homolog (mouse)	Hs.155968	ZFP103	2p11.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
131	-4.05	0.000155	0.565	1.142	162666	kininogen	Hs.77741	KNG	3q27
132	-4.05	0.000156	0.353	0.729	168282	group-specific component (vitamin D binding protein)	Hs.198246	GC	4q12-q13
133	-4.05	0.000157	0.678	0.841	168476	nucleoporin 88kDa	Hs.172108	NUP88	17p13.2
134	-4.04	0.000161	0.66	1.011	167801	Sec23 homolog A (S. cerevisiae)	Hs.272927	SEC23A	14q13.2
135	-4.01	0.00018	0.624	0.786	165731	tumor protein D52-like 1	Hs.16611	TPD52L1	6q22-q23
136	-4.01	0.000177	0.586	0.97	169253	paraoxonase 3	Hs.335322	PON3	7q21.3
137	-4.01	0.000179	0.841	1.036	159850	Homo sapiens cDNA FLJ34315 fis, clone FEBRA2008341, mRNA sequence	Hs.376655		14
138	-4	0.000182	0.69	1.057	167281	cell division cycle 2-like 5 (cholinesterase-related cell division controller)	Hs.59498	CDC2L5	7p13
139	-4	0.000185	0.589	0.913	165590	translocation protein 1	Hs.8146	TLOC1	3q26.2-q27
140	-3.99	0.00019	0.69	0.939	162599	haptoglobin	Hs.75990	HP	16q22.1
141	-3.97	0.000202	0.79	0.997	164028	ESTs, Weakly similar to ATDA_HUMAN Diamine acetyltransferase (Spermidine/spermine N(1)-acetyltransferase) (SSAT) (Putrescine acetyltransferase) [H.sapiens]	Hs.356269		X

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
142	-3.97	0.000205	0.418	0.896	166007	tyrosine aminotransferase	Hs.161640	TAT	16q22.1
143	-3.95	0.000219	0.828	1.188	165559	c-mer proto-oncogene tyrosine kinase	Hs.306178	MERTK	2q14.1
144	-3.95	0.000221	0.816	1.224	165133	basic leucine zipper and W2 domains 1	Hs.155291	BZW1	2q33
145	-3.94	0.000223	0.334	0.522	167542	KIAA0062 protein	Hs.89868	KIAA0062	8p21.2
146	-3.93	0.00023	0.504	0.902	169449	arginase, liver	166337 (IncytePD)		
147	-3.93	0.000231	0.649	0.78	167543	coagulation factor VIII, procoagulant component (hemophilia A)	Hs.79345	F8	Xq28
148	-3.93	0.000235	0.491	0.61	163368	CDw92 antigen	Hs.179902	CDW92	9q31.2
149	-3.91	0.000244	1.059	1.761	168931	heat shock 105kD	Hs.36927	HSP105B	13q12.2
150	-3.91	0.000245	0.406	0.687	165009	orosomucoid 1	Hs.572	ORM1	9q31-q32
151	-3.89	0.000264	0.37	0.662	162162	complement component 8, alpha polypeptide	Hs.93210	C8A	1p32
152	-3.89	0.000265	0.746	1.159	166110	2,4-dienoyl CoA reductase 1, mitochondrial	Hs.81548	DECR1	8q21.3
153	-3.88	0.000277	0.749	0.985	161689	growth hormone receptor	Hs.125180	GHR	5p13-p12
154	-3.87	0.000282	0.899	1.223	167617	selenoprotein P, plasma, 1	Hs.275775	SEPP1	5q31
155	-3.86	0.000291	0.644	0.938	161484	cytochrome P450,	Hs.106242	CYP4F3	19p13.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
						subfamily IVF, polypeptide 3 (leukotriene B4 omega hydroxylase)			
156	-3.85	0.000298	0.91	1.172	167551	microtubule-associated protein 7	Hs.146388	MAP7	6q23.2
157	-3.85	0.000299	0.604	0.895	169703	phosphoglucomutase 1	Hs.1869	PGM1	1p31
158	-3.85	0.000305	0.673	0.909	163040	Incyte EST	2593385 (IncytePD)		
159	-3.84	0.000311	0.602	0.807	165566	L-3-hydroxyacyl-Coenzyme A dehydrogenase, short chain	1550727 (IncytePD)		
160	-3.83	0.000322	1.017	1.192	162707	Homo sapiens clone 25038 mRNA sequence	Hs.306359		15
161	-3.83	0.000322	0.423	0.652	166674	paired basic amino acid cleaving system 4	Hs.170414	PACE4	15q26
162	-3.82	0.000327	0.732	1.378	165737	fatty acid binding protein 1, liver	Hs.380135	FABP1	2p11
163	-3.82	0.000334	0.596	0.86	168366	sterol carrier protein 2	Hs.75760	SCP2	1p32
164	-3.82	0.000334	0.809	1.044	165115	aconitase 1, soluble	Hs.154721	ACO1	9p22-p13
165	-3.82	0.000389	0.718	1.152	161732	plexin B1	Hs.278311	PLXNB1	3p21.31
166	-3.8	0.000349	0.854	1.28	162202	transferrin	Hs.396489	TF	3q21
167	-3.79	0.000361	0.553	0.886	167991	hydroxysteroid (17-beta) dehydrogenase 4	Hs.75441	HSD17B4	5q21

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
168	-3.79	0.000365	0.662	0.953	169717	progesterone receptor membrane component 1	Hs.90061	PGRMC1	Xq22-q24
169	-3.79	0.000367	0.554	1.088	165457	solute carrier family 27 (fatty acid transporter), member 2	Hs.11729	SLC27A2	15q21.2
170	-3.77	0.000389	0.687	1.101	164532	catalase	Hs.395771	CAT	11p13
171	-3.77	0.000401	0.969	1.28	162934	leucine carboxyl methyltransferase	Hs.8054	LCMT	16p12.3-16p12.1
172	-3.77	0.000391	0.583	0.822	160051	lymphocyte cytosolic protein 1 (L-plastin)	Hs.381099	LCP1	13q14.3
173	-3.77	0.000394	0.701	0.97	168394	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit	Hs.146812	HADHB	2p23
174	-3.75	0.000411	0.964	1.164	162323	EST	Hs.426542		4
175	-3.75	0.000419	0.689	1.067	160471	translational inhibitor protein p14.5	Hs.18426	UK114	8q22
176	-3.75	0.00042	0.624	0.823	163224	DC2 protein	Hs.103180	DC2	4q25
177	-3.73	0.000444	0.998	1.308	162773	calcium channel, voltage-dependent, beta 2 subunit	Hs.30941	CACNB2	10p12
178	-3.73	0.000454	0.88	1.1	166579	interleukin 18 receptor 1	Hs.159301	IL18R1	2q12
179	-3.72	0.00046	0.665	1.113	161872	serine (or cysteine) proteinase inhibitor, clade	Hs.76838	SERPINA7	Xq22.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
						A (alpha-1 antiproteinase, antitrypsin), member 7			
180	-3.71	0.000467	0.659	1.159	162012	lipoprotein, Lp(a)	Hs.119520	LPA	6q26-q27
181	-3.71	0.000469	0.859	1.179	163509	Hermansky-Pudlak syndrome 3	Hs.282804	HPS3	3q24
182	-3.68	0.000532	0.523	0.732	165011	tyrosylprotein sulfotransferase 1	Hs.421194	TPST1	7q11.21
183	-3.65	0.000577	0.649	0.875	164314	KIAA1450 protein	Hs.83243	KIAA1450	4q32.1
184	-3.64	0.000582	0.935	1.054	162882	RAB3A interacting protein (rabin3)-like 1	Hs.13759	RAB3IL1	11q12-q13.1
185	-3.62	0.000636	0.769	1.162	165530	cytochrome P450, subfamily IIJ (arachidonic acid epoxygenase) polypeptide 2	Hs.152096	CYP2J2	1p31.3-p31.2
186	-3.59	0.000679	0.487	0.924	166057	POU domain, class 1, transcription factor 1 (Pit1, growth hormone factor 1)	Hs.89394	POU1F1	3p11
187	-3.59	0.000703	0.95	1.228	167868	general transcription factor IIB	Hs.258561	GTF2B	1p22-p21
188	-3.58	0.000706	0.942	1.096	167779	general transcription factor IIE, polypeptide 2, beta 34kDa	Hs.77100	GTF2E2	8p21-p12
189	-3.58	0.000727	0.947	1.225	165329	Rab9 effector p40	Hs.19012	RAB9P40	9q34.11
190	-3.57	0.000735	0.62	1.11	166857	plasminogen	Hs.75576	PLG	6q26

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
191	-3.55	0.000775	0.838	1.215	165788	potassium inwardly-rectifying channel, subfamily J, member 8	Hs.102308	KCNJ8	12p11.23
192	-3.55	0.000778	0.662	1.001	167386	nicotinamide N-methyltransferase	604856 (IncytePD)		
193	-3.55	0.000795	0.671	0.802	163088	hypothetical protein FLJ21918	Hs.282093	FLJ21918	16q22.1
194	-3.55	0.00079	0.795	1.166	167385	electron-transfer-flavoprotein, alpha polypeptide (glutaric aciduria II)	Hs.169919	ETFA	15q23-q25
195	-3.54	0.000799	1.068	1.459	169569	spermidine/spermine N1-acetyltransferase	Hs.28491	SAT	Xp22.1
196	-3.54	0.000812	1.04	1.362	160982	ras responsive element binding protein 1	Hs.171942	RREB1	6p25
197	-3.53	0.00083	0.756	0.967	166818	tropomodulin	Hs.374849	TMOD	9q22.3
198	-3.52	0.000844	0.79	1.057	164368	Similar to RIKEN cDNA 1810013D05 gene [Homo sapiens], mRNA sequence	Hs.32699		12
199	-3.52	0.000848	0.609	1.001	160667	sorbitol dehydrogenase	Hs.878	SORD	15q15.3
200	-3.52	0.000851	0.713	0.894	160956	hypothetical protein A-211C6.1	Hs.28607	LOC57149	16p11.2
201	-3.52	0.000858	0.625	0.933	166778	phosphoenolpyruvate carboxykinase 2 (mitochondrial)	Hs.75812	PCK2	14q11.2

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
202	-3.52	0.000859	0.958	1.507	167552	lysosomal-associated membrane protein 2	Hs.8262	LAMP2	Xq24
203	-3.51	0.000891	1.012	1.281	160125	tumor protein, translationally-controlled 1	Hs.401448	TPT1	13q12-q14
204	-3.5	0.000901	0.991	1.197	161606	Fc fragment of IgG, receptor, transporter, alpha	Hs.111903	FCGRT	19q13.3
205	-3.5	0.000914	1.005	1.238	165593	transmembrane 7 superfamily member 1 (upregulated in kidney)	Hs.15791	TM7SF1	1q42-q43
206	-3.5	0.000915	1.009	1.267	160129	MAP/microtubule affinity-regulating kinase 2	Hs.157199	MARK2	11q12-q13
207	-3.47	0.000997	0.457	0.74	168320	lactate dehydrogenase A	Hs.2795	LDHA	11p15.4
208	3.47	0.000996	1.098	0.826	160605	P311 protein	Hs.142827	P311	5q22.1
209	3.48	0.000971	0.938	0.81	165174	Homo sapiens cDNA FLJ35787 fis, clone TESTI2005672, highly similar to UBIQUINOL-CYTOCHROME C REDUCTASE COMPLEX CORE PROTEIN 2 PRECURSOR (EC 1.10.2.2), mRNA sequence	Hs.265591		16
210	3.49	0.000953	1.083	0.92	166833	solute carrier family 17 (anion/sugar transporter), member 5	Hs.117865	SLC17A5	6q14-q15

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
211	3.53	0.000822	1.111	0.953	165413	WIRE protein	Hs.13996	WIRE	17q21.1
212	3.56	0.000759	1.08	0.941	166348	epidermal growth factor receptor pathway substrate 8-related protein 1	Hs.28907	EPS8R1	19q13.42
213	3.57	0.000725	1.048	0.9	163115	ESTs, Moderately similar to hypothetical protein FLJ20234 [Homo sapiens] [H.sapiens]	Hs.119629		14
214	3.59	0.000684	1.124	0.963	163579	ESTs	Hs.194441		6
215	3.59	0.00068	1.835	1.306	161090	KIAA1641 protein	Hs.44566	KIAA1641	2q11.1
216	3.6	0.000688	1.173	0.981	161354	p21/Cdc42/Rac1-activated kinase 1 (STE20 homolog, yeast)	Hs.64056	PAK1	11q13-q14
217	3.61	0.000661	0.947	0.82	162677	Human BRCA2 region, mRNA sequence CG011	Hs.142907		13
218	3.64	0.000582	0.853	0.653	161085	polymerase (DNA directed), delta 1, catalytic subunit 125kDa	Hs.99890	POLD1	19q13.3
219	3.65	0.000572	1.141	0.985	161518	H2A histone family, member A	Hs.121017	H2AFA	6p22.2-p21.1
220	3.65	0.000571	1.232	1.062	163109	mitochondrial ribosomal protein L43	Hs.151945	MRPL43	10q24.1-q24.3
221	3.67	0.000537	1.014	0.841	164845	thioredoxin domain containing 4 (endoplasmic reticulum)	Hs.154023	TXNDC4	9q22.33

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
222	3.67	0.000538	0.677	0.528	162564	A kinase (PRKA) anchor protein (yotiao) 9	Hs.58103	AKAP9	7q21-q22
223	3.68	0.000522	1.301	1.059	164727	ESTs	Hs.125038		8
224	3.68	0.00052	0.898	0.781	161620	H4 histone family, member A [Homo sapiens], mRNA sequence	Hs.278483		3
225	3.7	0.000484	0.976	0.854	161334	hypothetical protein 20D7-FC4	Hs.128702	20D7-FC4	19q13.3
226	3.74	0.000428	1.081	0.871	163536	transducer of ERBB2, 2	Hs.4994	TOB2	22q13.2-q13.31
227	3.77	0.000411	1.376	1.018	162152	claudin 4	Hs.5372	CLDN4	7q11.23
228	3.83	0.00033	1.138	0.936	169742	ESTs, Moderately similar to hypothetical protein FLJ20378 [Homo sapiens] [H.sapiens]	Hs.143992		2
229	3.84	0.000311	1.234	1.035	161058	multiple endocrine neoplasia I	Hs.423348	MEN1	11q13
230	3.84	0.000311	0.765	0.619	161813	KIAA0874 protein	Hs.27973	KIAA0874	18p11.21
231	3.84	0.000311	1.227	1.01	168511	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	Hs.78934	MSH2	2p22-p21
232	3.84	0.000309	1.265	1.031	161873	Incyte EST	3031912 (IncytePD)		
233	3.89	0.000263	1.174	1.002	169310	nucleoporin 62kDa	Hs.9877	NUP62	19q13.33

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
234	3.9	0.000259	2.07	1.521	168933	midkine (neurite growth-promoting factor 2)	Hs.82045	MDK	11p11.2
235	3.96	0.000213	1.058	0.893	163495	hypothetical protein FLJ11280	Hs.3346	FLJ11280	1q21.2
236	3.96	0.000209	0.749	0.557	168500	Homo sapiens cDNA: FLJ21930 fis, clone HEP04301, highly similar to HSU90916 Human clone 23815 mRNA sequence	Hs.82845		11
237	3.96	0.000207	1.258	0.984	168246	thyroid hormone receptor interactor 13	Hs.6566	TRIP13	5p15.33
238	3.98	0.000199	1.087	0.914	164713	Homo sapiens full length insert cDNA clone ZC18H06, mRNA sequence	Hs.384561		19
239	4.03	0.000168	1.324	0.862	169559	E74-like factor 3 (ets domain transcription factor, epithelial-specific)	Hs.166096	ELF3	1q32.2
240	4.04	0.000162	1.138	0.906	164262	membrane protein, palmitoylated 6 (MAGUK p55 subfamily member 6)	Hs.108931	MPP6	7p15
241	4.04	0.000164	1.187	0.998	161661	hypothetical protein FLJ10520	Hs.77510	FLJ10520	16q22.3
242	4.05	0.000159	1.15	0.926	163071	Homo sapiens cDNA: FLJ21409 fis, clone COL03924, mRNA sequence	Hs.172129		5

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
243	4.12	0.000126	1.058	0.922	165465	KIAA0195 gene product	Hs.301132	KIAA0195	17q25.2
244	4.14	0.000117	1.265	1.023	164085	ESTs	Hs.107845		2
245	4.14	0.000119	1.301	1.049	166229	hypothetical protein FLJ11362	Hs.8929	FLJ11362	Xq25-q26.1
246	4.18	0.000102	1.027	0.872	166228	huntingtin (Huntington disease)	Hs.79391	HD	4p16.3
247	4.21	9.10E-05	0.614	0.427	169583	neurogranin (protein kinase C substrate, RC3)	Hs.26944	NRGN	11q24
248	4.3	6.80E-05	1.37	1.01	160913	claudin 4	Hs.5372	CLDN4	7q11.23
249	4.31	6.60E-05	1.063	0.844	168965	formin binding protein 3	Hs.107213	FNBP3	2q23.3
250	4.35	5.80E-05	1.154	0.878	166849	p53-responsive gene 5	1510581 (IncytePD)		
251	4.37	5.30E-05	1.021	0.816	167919	KIAA1361 protein	Hs.15119	KIAA1361	17q11.1
252	4.45	4.00E-05	1.219	0.977	166837	ESTs	Hs.279482		2
253	4.45	4.10E-05	1.278	0.974	168977	Homo sapiens cDNA FLJ34031 fis, clone FCBBF2003895, mRNA sequence	Hs.340316		19
254	4.49	3.50E-05	1.06	0.93	166408	hypothetical protein FLJ39514	Hs.48565	FLJ39514	4q11
255	4.49	3.50E-05	1.233	0.952	167009	protein kinase C, iota	Hs.1904	PRKCI	3q26.3
256	4.6	2.40E-05	1.237	1.011	168029	small nuclear	Hs.173255	SNRPA	19q13.1

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
						ribonucleoprotein polypeptide A			
257	4.61	2.40E-05	0.838	0.632	169587	v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog	Hs.433714	KRAS2	12p12.1
258	4.61	2.40E-05	1.425	0.974	163235	FLJ00005 protein	Hs.367690	FLJ00005	15q22.33
259	4.63	2.20E-05	1.153	0.953	161066	hypothetical protein from clone 24796	Hs.27191	LOC57146	16p12
260	4.67	1.90E-05	0.967	0.784	165515	3-phosphoinositide dependent protein kinase-1	Hs.154729	PDPK1	16p13.3
261	4.67	1.90E-05	1.035	0.775	169403	protein phosphatase 1, regulatory (inhibitor) subunit 12A	Hs.16533	PPP1R12A	12q15-q21
262	4.71	1.60E-05	1.14	0.951	169490	hypothetical protein DKFZp564K0322	Hs.97876	DKFZP564K032	19q13.32
263	4.71	1.60E-05	3.6	1.727	160089	tumor-associated calcium signal transducer 1	Hs.692	TACSTD1	2p21
264	4.73	1.50E-05	1.055	0.889	169508	ATPase, Cu++ transporting, alpha polypeptide (Menkes syndrome)	Hs.606	ATP7A	Xq13.2-q13.3
265	4.8	1.20E-05	1.478	1.137	163214	hypothetical protein FLJ22548 similar to gene trap PAT 12	Hs.103267	FLJ22548	12q14.3
266	5.16	3.00E-06	1.12	0.89	168509	ESTs, Weakly similar to	Hs.99398		14

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
						KHLX_HUMAN Kelch-like protein X [H.sapiens]			
267	5.17	3.00E-06	0.855	0.668	166434	hypothetical protein FLJ13213	Hs.331328	FLJ13213	15q21.2
268	5.37	1.00E-06	1.164	0.929	161233	Incyte EST	1602194 (IncytePD)		
269	5.55	p < 0.000001	1.449	0.963	167498	protocadherin 17	Hs.106511	PCDH17	13q14.3
270	5.99	p < 0.000001	1.201	0.896	160943	Homo sapiens clone 24630 mRNA sequence	Hs.171553		3
271	6.36	p < 0.000001	1.345	1.012	165379	hypothetical protein BC008647	Hs.102480	LOC91875	14q11.1
272	6.36	p < 0.000001	1.376	0.962	167992	KIAA1557 protein	Hs.6185	KIAA1557	12p11.21
273	6.37	p < 0.000001	1.229	0.824	166068	ectodermal-neural cortex (with BTB-like domain)	Hs.104925	ENC1	5q12-q13.3

[0258] The top 25 genes with the lowest parametric p-values ($p<0.000001$) were selected from the 273-gene set and this set gave rise to a comparable result as the 273-gene set. These 25 genes significant for indicating a liver disease patient's risk of developing HCC, their gene symbols, their map positions, and their UG Cluster identifiers are presented in Table 6. A further set of 10 significant genes for predicting the risk of developing HCC in a patient suffering from a severe liver disease has been determined in a similar manner and is presented in Table 7.

Table 6. 25 significant genes for identifying patients likely to develop HCC by the compound covariate predictor analysis and their values necessary for computing multifactorial L value in the prediction model.

	t-value	Parametric p-value	% CV support	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
1	-7.28	0.0000001	100	0.603	0.903	160198	cofilin 2 (muscle)	Hs.180141	CFL2	14q
2	-6.53	0.0000001	100	0.985	1.607	168023	Fc fragment of IgG, high affinity Ia, receptor for (CD64)	Hs.77424	FCGR1A	1q21.2-q21.3
3	-6.46	0.0000001	100	0.643	1.175	162315	calcium channel, voltage-dependent, beta 3 subunit	Hs.250712	CACNB3	12q13
4	-6.18	0.0000001	100	0.688	1.112	160302	myosin IB	Hs.121576	MYO1B	2q12-q34
5	-6.16	0.0000001	100	0.473	1.161	169417	ceruloplasmin (ferroxidase)	Hs.296634	CP	3q23-q25
6	-6.1	0.0000001	100	0.876	1.18	161756	albumin	Hs.184411	ALB	4q11-q13
7	-6.04	0.0000001	100	0.719	1.224	162290	UDP-N-acteylglucosamine pyrophosphorylase 1	Hs.21293	UAP1	1q23.1
8	-6.01	0.0000001	100	0.534	1.141	162538	Unknown [Homo sapiens], mRNA sequence	Hs.367982		16
9	-5.94	0.0000001	100	0.491	0.714	168634	chromosome 20 open reading frame 3	Hs.22391	C20orf3	20p11.22-p11.21
10	-5.93	0.0000001	100	0.756	1.276	164136	acyl-Coenzyme A dehydrogenase,	Hs.1209	ACADL	2q34-q35

	t-value	Parametric p-value	% CV support	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
							long chain			
11	-5.9	0.0000001	100	0.864	1.181	163874	KIAA0092 gene product	Hs.151791	KIAA0092	11q21
12	-5.88	0.0000001	100	0.728	0.925	163096	CGI-26 protein	Hs.24332	CGI-26	12p12.3
13	-5.73	0.0000001	100	0.616	1.133	160233	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 3	Hs.38018	DYRK3	1q32
14	-5.67	0.0000001	100	0.786	1.071	160436	Similar to hypothetical protein PRO2831 [Homo sapiens], mRNA sequence	Hs.406646		15
15	-5.65	0.0000001	100	0.761	1.382	160795	hepatic leukemia factor	Hs.433707	HLF	17q22
16	-5.61	0.0000001	100	0.314	0.798	161944	complement component 9	Hs.1290	C9	5p14-p12
17	-5.6	0.0000001	100	0.506	0.703	167718	ATP-binding cassette, sub-family A (ABC1), member 1	Hs.211562	ABCA1	9q31.1
18	-5.58	0.0000001	100	0.65	0.912	168437	KIAA0843 protein	Hs.26777	KIAA0843	5q32
19	-5.57	0.0000001	100	0.843	1.087	162884	intracellular membrane-associated calcium-independent phospholipase A2 gamma	Hs.44198	IPLA2(GAMMA)	7q31

	t-value	Parametric p-value	% CV support	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	Unique id	Description	UG cluster	Gene symbol	Map
20	-5.48	0.0000001	100	0.657	1.065	166910	SIPL protein	Hs.64322	SIPL	2p25.3
269	5.55	0.0000001	100	1.449	0.963	167498	protocadherin 17	Hs.106511	PCDH17	13q14.3
270	5.99	0.0000001	100	1.201	0.896	160943	Homo sapiens clone 24630 mRNA sequence	Hs.171553		3
271	6.36	0.0000001	100	1.345	1.012	165379	hypothetical protein BC008647	Hs.102480	LOC91875	14q11.1
272	6.36	0.0000001	100	1.376	0.962	167992	KIAA1557 protein	Hs.6185	KIAA1557	12p11.21
273	6.37	0.0000001	100	1.229	0.824	166068	ectodermal-neural cortex (with BTB-like domain)	Hs.104925	ENC1	5q12-q13.3

These 25 genes were selected by the 10 smallest parametric p values ($p<0.000001$).

Table 7. 10 Significant genes for predicting HCC development and their values necessary for computing multifactorial L value in the prediction model.

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	High/Low	Unique id	Description	UG cluster	Gene symbol	Map
103	-4.26	7.70E-05	0.406	0.89	0.45618	161362	tryptophan 2,3-dioxygenase	Hs.183671	TDO2	4q31-q32
77	-4.57	3.60E-05	0.421	0.88	0.478409	161748	acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)	Hs.37	ACAT1	11q22.3-q23.1
42	-5.02	5.00E-06	0.506	1.137	0.445031	162311	serine (or cysteine) proteinase inhibitor, clade C (antithrombin), member 1	Hs.75599	SERPINC1	1q23-q25.1
8	-6.01	p < 0.000001	0.534	1.141	0.468011	162538	Unknown [Homo sapiens], mRNA sequence	Hs.367982		16
63	-4.7	1.80E-05	0.355	0.917	0.387132	162617	FK506 binding protein 5	Hs.7557	FKBP5	6p21.3-21.2
131	-4.05	0.000155	0.565	1.142	0.494746	162666	kininogen	Hs.77741	KNG	3q27
121	-4.14	0.000116	0.59	1.176	0.501701	164863	solute carrier family 2 (facilitated glucose transporter), member 2	Hs.167584	SLC2A2	3q26.1-q26.2
142	-3.97	0.000205	0.418	0.896	0.466518	166007	tyrosine aminotransferase	Hs.161640	TAT	16q22.1
116	-4.19	9.90E-05	0.507	1.102	0.460073	169375	cytochrome P450, subfamily IIC (mephenytoin 4-	Hs.167529	CYP2C9	10q24

	t-value	Parametric p-value	Geom mean of ratios in class 1: High	Geom mean of ratios in class 2: Low	High/Low	Unique id	Description	UG cluster	Gene symbol	Map
							hydroxylase), polypeptide 9			
5	-6.16	p < 0.000001	0.473	1.161	0.407407	169417	ceruloplasmin (ferroxidase)	Hs.296634	CP	3q23-q25

WHAT IS CLAIMED IS:

1. A method for identifying potential therapeutic targets for inhibiting metastasis in a patient suffering from hepatocellular carcinoma (HCC), comprising the steps of:

a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a metastatic HCC patient;

b) capturing markers from the sample and generating a first signal;

c) repeating steps a) and b) with a sample from a non-metastatic HCC patient and thereby generating a second signal; and

d) comparing the first and second signals and thereby identifying a subset of cellular markers whose level is different in the first and second signals, wherein the subset of cellular markers are potential therapeutic targets for treating HCC metastasis in an HCC patient.

2. The method of claim 1, wherein a signal generated from a normal non-cancerous sample on an array identical to the array of step a) is subtracted in steps b) and c) to generate the first and second signals.

3. A method for predicting the metastatic potential in a patient suffering from hepatocellular carcinoma (HCC), comprising the steps of:

a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a metastatic HCC patient, the set of cellular markers comprising at least ten genes or proteins encoded by genes independently selected from the genes of Table 2;

b) capturing markers from the sample;

c) generating a first signal from the captured markers of step b);

d) repeating steps a) to c) with a sample from a non-metastatic HCC patient and thereby generating a second signal;

e) repeating steps a) to c) with a sample from an HCC patient with unknown metastatic potential and thereby generating a third signal; and

f) comparing the third signal to the first and the second signals and thereby determining the metastatic potential of the HCC patient of step e).

4 . The method of claim 3, wherein the set of cellular markers comprises at least 20 genes or proteins encoded by genes independently selected from the genes of Table 2.

5. The method of claim 4, wherein the set of cellular markers comprises at least 50 genes or proteins encoded by genes independently selected from the genes of Table 2.

6. The method of claim 5, wherein the set of cellular markers comprises at least 100 genes or proteins encoded by genes independently selected from the genes of Table 2.

7. The method of claim 6, wherein the set of cellular markers comprises the genes or proteins encoded by genes of Table 2.

8. The method of claim 3, wherein the set of cellular markers comprises the genes or proteins encoded by genes of Table 4.

9. The method of claim 3, wherein the set of cellular markers comprises the genes or proteins encoded by genes of Unigene numbers Hs.313, Hs.69707, Hs.222, Hs.63984, Hs.75573, Hs.177687, Hs.69707, Hs.222, Hs.323712, and Hs.63984.

10. The method of claim 3, wherein the sample of steps a) and b), the sample of step d), and the sample of step e) are liver tissue extracts.

11. The method of claim 3, wherein the array of step a) is a genomic array.

12. The method of claim 3, wherein the array of step a) is a proteomic array.

13. A method for identifying potential therapeutic targets for preventing hepatocellular carcinoma (HCC) in a patient suffering from a chronic liver disease, comprising the steps of:

a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a patient with a chronic liver disease and a high risk of developing HCC;

b) capturing markers from the sample and generating a first signal;

c) repeating steps a) and b) with a sample from a patient with a chronic liver disease and a low risk of developing HCC and thereby generating a second signal; and

d) comparing the first and second signals and thereby identifying a subset of cellular markers whose level is different in the first and second signals, wherein the subset of cellular markers are potential therapeutic targets for preventing HCC in a patient with a chronic liver disease.

14. The method of claim **13**, wherein a signal generated from a normal non-canerous sample on an array identical to the array of step a) is subtracted in steps b) and c) to generate the first and second signals.

15. A method for predicting the risk of developing hepatocellular carcinoma (HCC) in a patient suffering from a chronic liver disease, comprising the steps of:

a) contacting an array comprising capture reagents for a set of cellular markers with a sample from a patient with a chronic liver disease and a high risk of HCC, the set of cellular markers comprising at least ten genes or proteins encoded by genes independently selected from the genes of Table 5;

b) capturing markers from the sample;

c) generating a first signal from the captured markers of step b);

d) repeating steps a) to c) with a sample from a patient with a chronic liver disease and a low risk of HCC and thereby generating a second signal;

e) repeating steps a) to c) with a sample from a patient with a chronic liver disease and an unknown risk of HCC and thereby generating a third signal; and

f) comparing the third signal to the first and the second signals and thereby determining the risk of developing HCC in the patient of step e).

16. The method of claim **15**, wherein the set of cellular markers comprises at least 20 genes or proteins encoded by genes independently selected from the genes of Table 5.

17. The method of claim **16**, wherein the set of cellular markers comprises at least 50 genes or proteins encoded by genes independently selected from the genes of Table 5.

18. The method of claim 17, wherein the set of cellular markers comprises at least 100 genes or proteins encoded by genes independently selected from the genes of Table 5.

19. The method of claim 18, wherein the set of cellular markers comprises the genes or proteins encoded by genes of Table 5.

20. The method of claim 15, wherein the set of cellular markers comprises the genes or proteins encoded by genes of Table 6.

21. The method of claim 15, wherein the set of cellular markers comprises the genes or proteins encoded by genes of Table 7.

22. The method of claim 15, wherein the sample of steps a) and b), the sample of step d), and the sample of step e) are liver tissue extracts.

23. The method of claim 15, wherein the array of step a) is a genomic array.

24. The method of claim 15, wherein the array of step a) is a proteomic array.

25. The method of claim 15, wherein the patient of step a) suffers from a disease selected from the groups consisting of hepatitis B, hepatitis C, hemachromatosis, and Wilson's disease.

26. The method of claim 15, wherein the patient of step d) suffers from alcoholic liver disease, autoimmune hepatitis, or primary biliary cirrhosis.

27. The method of claim 15, wherein the patient of step e) suffers from a disease selected from the group consisting of hepatitis B, hepatitis C, hemochromatosis, Wilson's disease, alcoholic liver disease, autoimmune hepatitis, and primary biliary cirrhosis.

28. A computer readable medium comprising:

a) code for a first data set, derived from a first signal from an array comprising capture reagents for a set of cellular markers after contact with a sample from a metastatic HCC patient, the set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 2;

b) code for a second data set, derived from a second signal from an array identical to the array of a) after contact with a sample from a non-metastatic HCC patient;

c) code for a third data set, derived from a third signal from an array identical to the array of a) after contact with a sample from a HCC patient with unknown metastatic potential; and

d) code for comparing the third data set with the first and second data sets.

29. A digital computer comprising the computer readable medium of claim **28**.

30. A system comprising:

a) a digital computer of claim **29**;

b) a chip with an array comprising capture reagents for a set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 2; and

c) a reader capable of registering a signal from the array after contact with a sample.

31. A computer readable medium comprising:

a) code for a first data set, derived from a first signal from an array comprising capture reagents for a set of cellular markers after contact with a sample from a patient with a chronic liver disease and a high risk of HCC, the set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 5;

b) code for a second data set, derived from a second signal from an array identical to the array of a) after contact with a sample from a patient with a chronic liver disease and a low risk of HCC;

c) code for a third data set, derived from a third signal from an array identical to the array of a) after contact with a sample from a patient with a chronic liver disease and an unknown risk of HCC; and

d) code for comparing the third data set with the first and second data sets.

32. A digital computer comprising the computer readable medium of claim 31.

33. A system comprising:

a) a digital computer of claim 32;

b) a chip with an array comprising capture reagents for a set of cellular markers comprising at least 10 genes or proteins encoded by genes independently selected from the genes of Table 5; and

c) a reader capable of registering a signal from the array after contact with a sample.

34. A method for inhibiting hepatocellular carcinoma (HCC) metastasis in a patient suffering from HCC, the method comprising the step of suppressing osteopontin (OPN) activity.

35. The method of claim 34, wherein the step of suppressing osteopontin (OPN) activity is accomplished by inhibiting OPN expression.

36. The method of claim 35, wherein an antisense polynucleotide is used to inhibit OPN expression.

37. The method of claim 34, wherein the step of suppressing osteopontin (OPN) activity is accomplished by inhibiting the specific binding between OPN and OPN receptor.

38. The method of claim 37, wherein an OPN antagonist is used to inhibit the specific binding between OPN and OPN receptor.

39. The method of claim 37, wherein an anti-OPN antibody is used to inhibit the specific binding between OPN and OPN receptor.

40. A method for inhibiting the development of hepatocellular carcinoma (HCC) in a patient suffering from a chronic liver disease, comprising the step of suppressing EpCAM activity.

41. The method of claim 40, wherein the step of suppressing EpCAM activity is accomplished by inhibiting EpCAM expression.

42. The method of claim **41**, wherein an antisense polynucleotide is used to inhibit EpCAM expression.

43. The method of claim **41**, wherein a small inhibitory RNA is used to inhibit EpCAM expression.

44. The method of claim **40**, wherein the step of suppressing EpCAM activity is accomplished by inhibiting the specific binding between EpCAM and EpCAM receptor.

45. The method of claim **44**, wherein an anti-EpCAM antibody is used to inhibit the specific binding between EpCAM and EpCAM receptor.

Figure 1

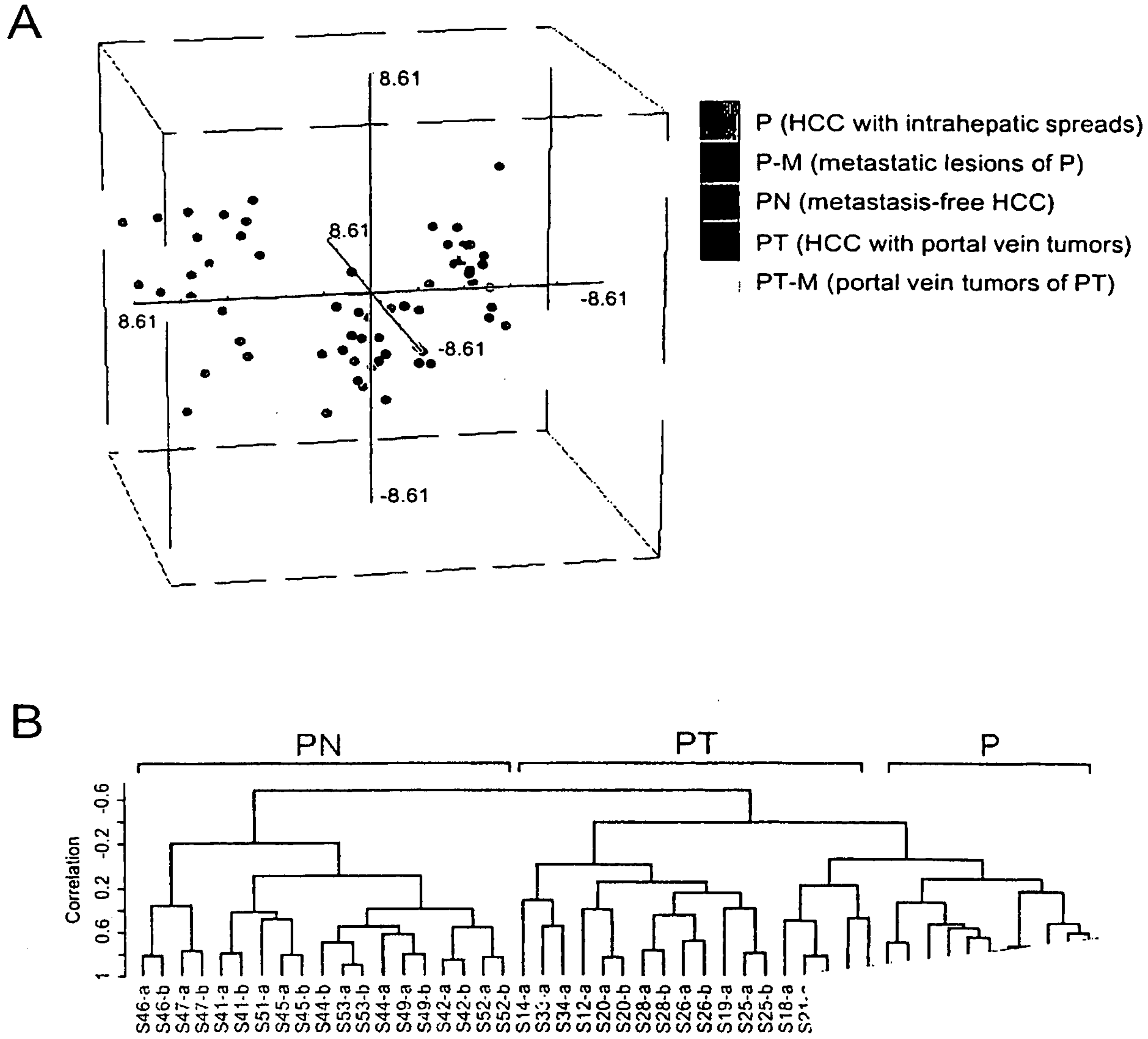

A
8.61
8.61
8.61
-8.61
-8.61
-8.61
P (HCC with intrahepatic spreads)
P-M (metastatic lesions of P)
PN (metastasis-free HCC)
PT (HCC with portal vein tumors)
PT-M (portal vein tumors of PT)
B
PN
PT
P
Correlation
-0.6
-0.2
0.2
0.6
1
S46-a
S46-b
S47-a
S47-b
S41-a
S41-b
S51-a
S45-a
S45-b
S44-b
S53-a
S53-b
S44-a
S49-a
S49-b
S42-a
S42-b
S52-a
S52-b
S14-a
S33-a
S34-a
S12-a
S20-a
S20-b
S28-a
S28-b
S26-a
S26-b
S19-a
S25-a
S25-b
S18-a

Figure 2

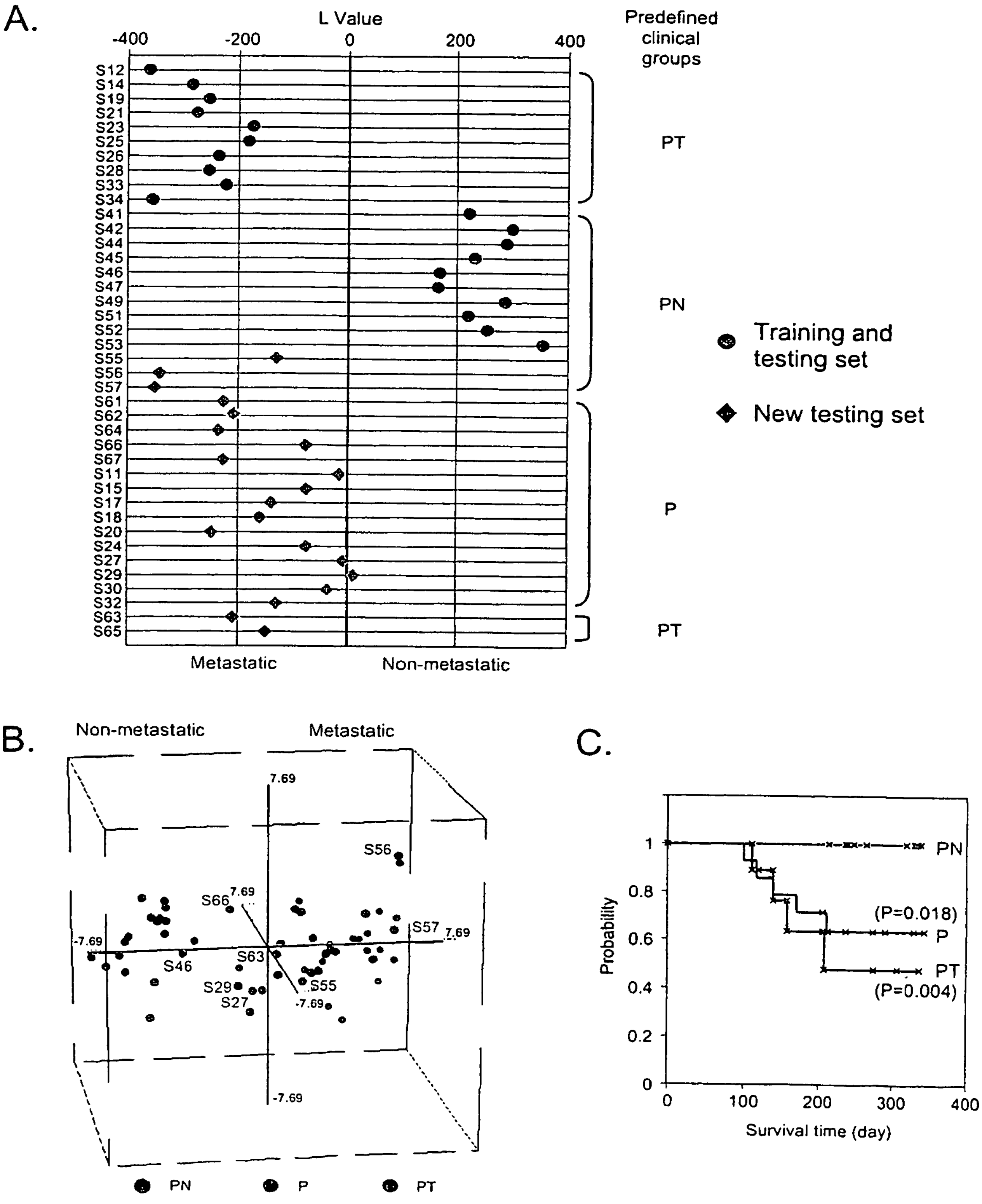
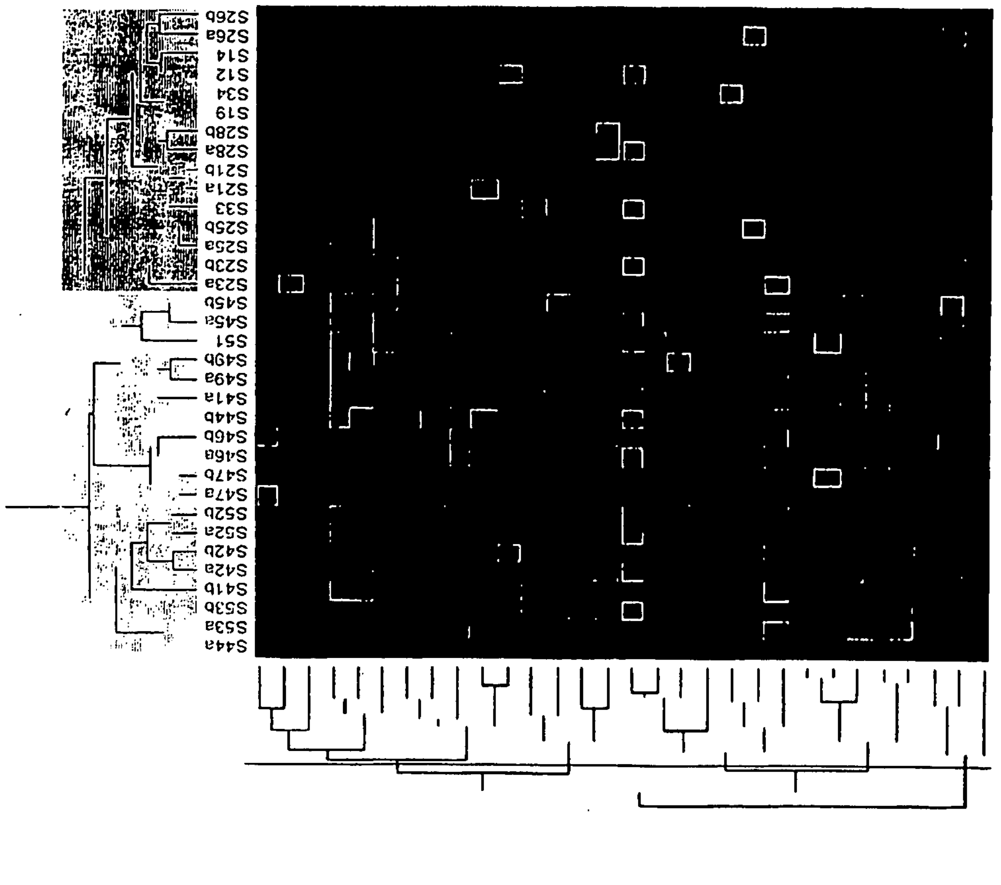

A.
L Value
-400
-200
0
200
400
Predefined clinical groups
S12
S14
S19
S21
S23
S25
S26
S28
S33
S34
S41
S42
S44
S45
S46
S47
S49
S51
S52
S53
S55
S56
S57
S61
S62
S64
S66
S67
S11
S15
S17
S18
S20
S24
S27
S29
S30
S32
S63
S65
PT
PN
P
PT
Training and testing set
New testing set
Metastatic
Non-metastatic
B.
Non-metastatic
Metastatic
7.69
-7.69
S56
S66
S57
S46
S63
S29
S27
S55
PN
P
PT
C.
Probability
1
0.8
0.6
0.4
0.2
0
PN
(P=0.018)
P
PT
(P=0.004)
0
100
200
300
400
Survival time (day)

Figure 3A

1/8
1/4
1/2
1
2
4
8



Symbol	UG cluster	Description
MST4	Hs.23643	serine/threonine protein kinase MASK
KHK	Hs.81454	ketohexokinase (fructokinase)
CYP4F12	Hs.180570	cytochrome P450 isoform 4F12
AKR1C4	Hs.177687	aldo-keto reductase family 1, member C4
CES1	Hs.76688	carboxylesterase 1
OPN	Hs.313	Osteopontin
GPRK5	Hs.211569	G protein-coupled receptor kinase 5
HCGII-7	Hs.69707	HCGII-7 protein
ENO3	Hs.118804	enolase 3, (beta, muscle)
		Unknown
CAT56	Hs.118354	CAT56 protein
NR1D1	Hs.276916	nuclear receptor subfamily 1, group D, member 1
FZD2	Hs.81217	frizzled (Drosophila) homolog 2
CENPE	Hs.75573	centromere protein E (312kD)
ITGA9	Hs.222	integrin, alpha 9
MMP9	Hs.151738	matrix metalloproteinase 9
SERPINB5	Hs.55279	serine (or cysteine) proteinase inhibitor, member 5
YES1	Hs.194148	v-yes-1 homolog 1
INPP5B	Hs.182577	inositol polyphosphate-5-phosphatase
IL2RB	Hs.75596	interleukin 2 receptor, beta
		Unknown
LRP6	Hs.23672	low density lipoprotein receptor-related protein 6
CD37	Hs.153053	CD37 antigen
MDFI	Hs.153203	MyoD family inhibitor
IRF2	Hs.83795	interferon regulatory factor 2
ASPH	Hs.283664	aspartate beta-hydroxylase
LILRA2	Hs.94498	leukocyte immunoglobulin-like receptor
RAB28	Hs.296371	RAB28, member RAS oncogene family
IGFBP6	Hs.274313	insulin-like growth factor binding protein 6
TYMSTR	Hs.34526	G protein-coupled receptor

Figure 3

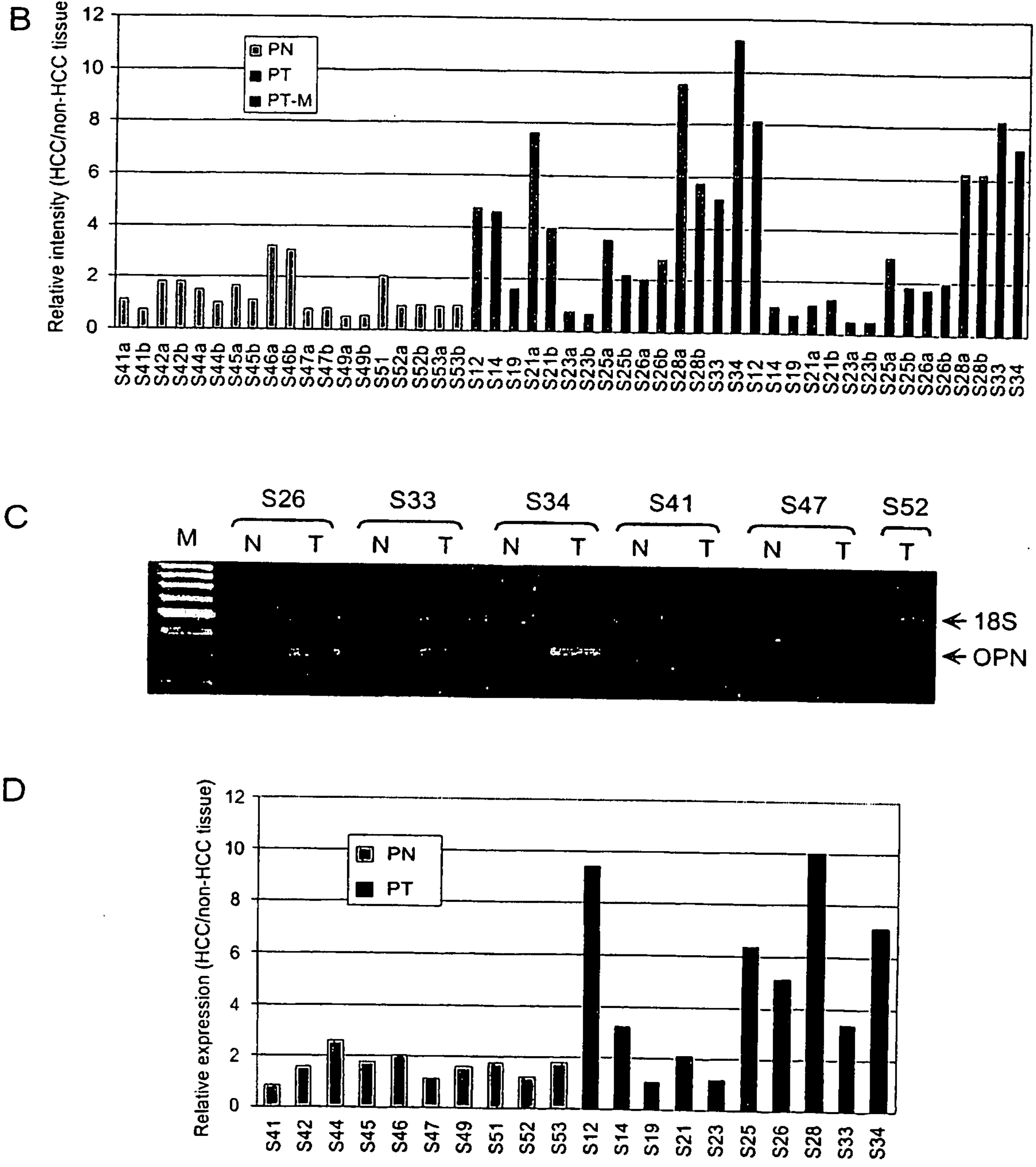

B
Relative intensity (HCC/non-HCC tissue)
12
10
8
6
4
2
0
PN
PT
PT-M
S41a
S41b
S42a
S42b
S44a
S44b
S45a
S45b
S46a
S46b
S47a
S47b
S49a
S49b
S51
S52a
S52b
S53a
S53b
S12
S14
S19
S21a
S21b
S23a
S23b
S25a
S25b
S26a
S26b
S28a
S28b
S33
S34
S12
S14
S19
S21a
S21b
S23a
S23b
S25a
S25b
S26a
S26b
S28a
S28b
S33
S34
C
S26
S33
S34
S41
S47
S52
M
N
T
N
T
N
T
N
T
N
T
T
← 18S
← OPN
D
Relative expression (HCC/non-HCC tissue)
12
10
8
6
4
2
0
PN
PT
S41
S42
S44
S45
S46
S47
S49
S51
S52
S53
S12
S14
S19
S21
S23
S25
S26
S28
S33
S34

Figure 4

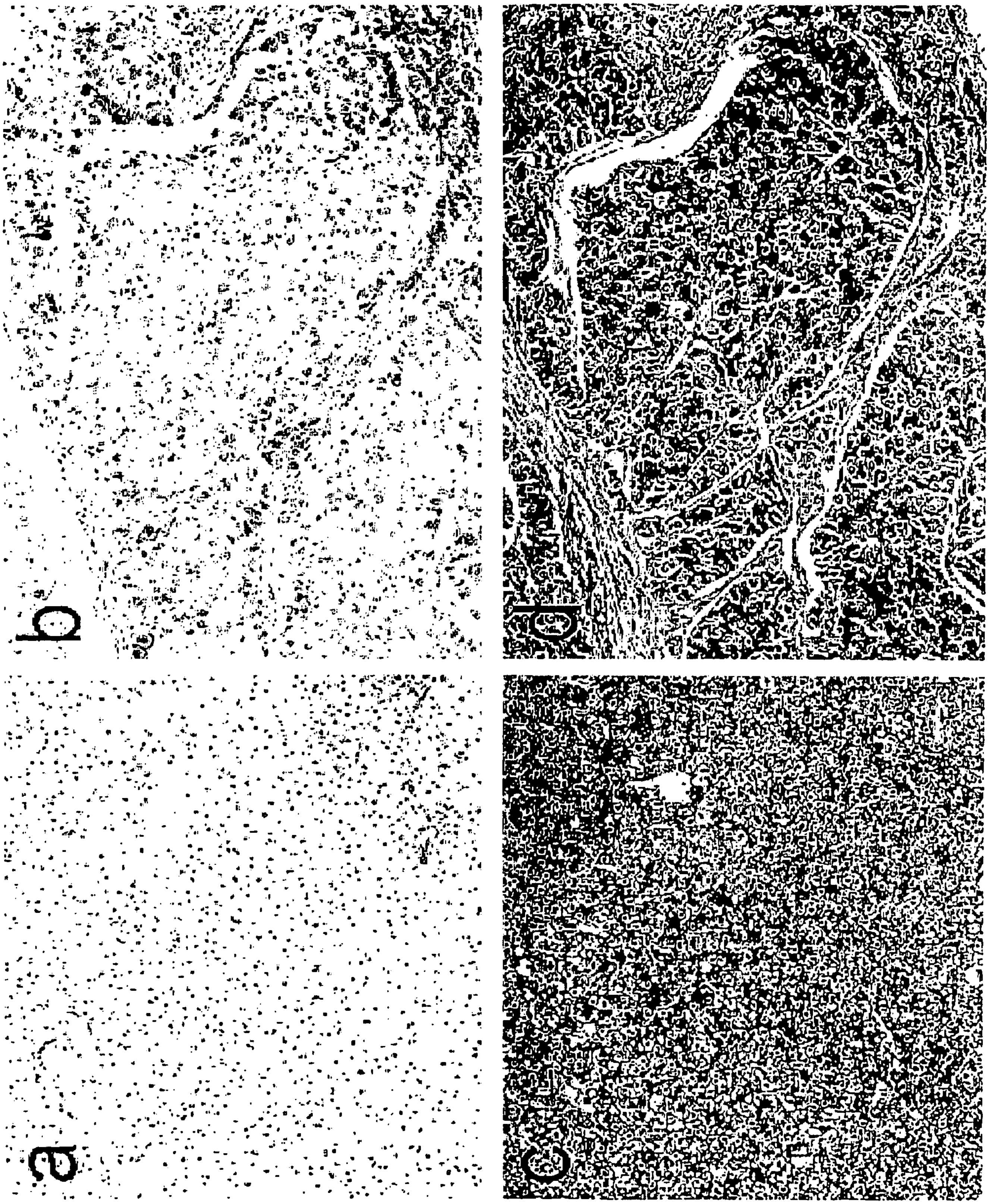


Figure 5

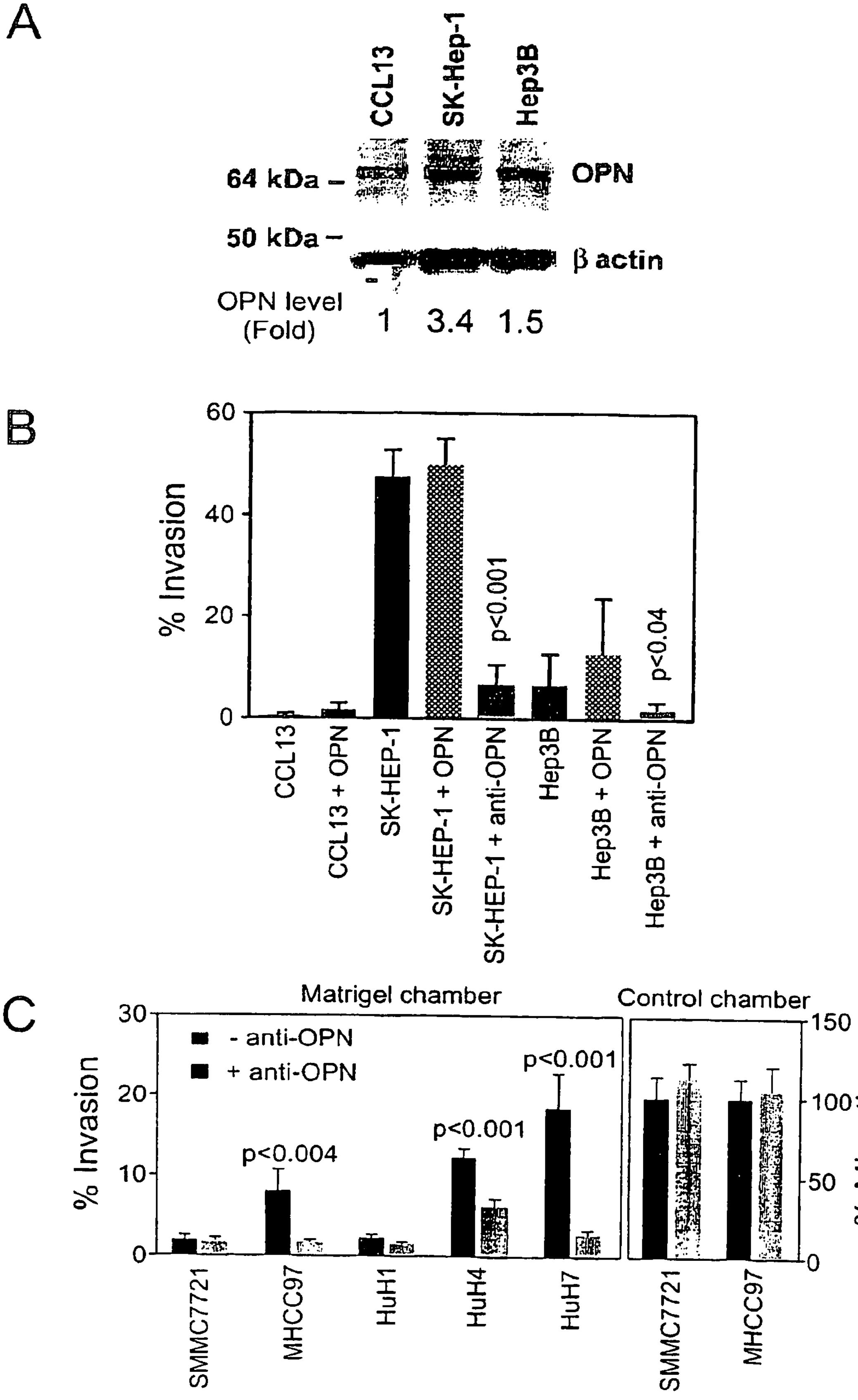

A
CCL13
SK-Hep-1
Hep3B
64 kDa –
OPN
50 kDa –
β actin
OPN level
(Fold)
1
3.4
1.5
B
% Invasion
60
40
20
0
p<0.001
p<0.04
CCL13
CCL13 + OPN
SK-HEP-1
SK-HEP-1 + OPN
SK-HEP-1 + anti-OPN
Hep3B
Hep3B + OPN
Hep3B + anti-OPN
C
Matrigel chamber
Control chamber
■ - anti-OPN
■ + anti-OPN
% Invasion
30
20
10
0
p<0.004
p<0.001
p<0.001
% Migration
150
100
50
0
SMMC7721
MHCC97
HuH1
HuH4
HuH7
SMMC7721
MHCC97

Figure 5

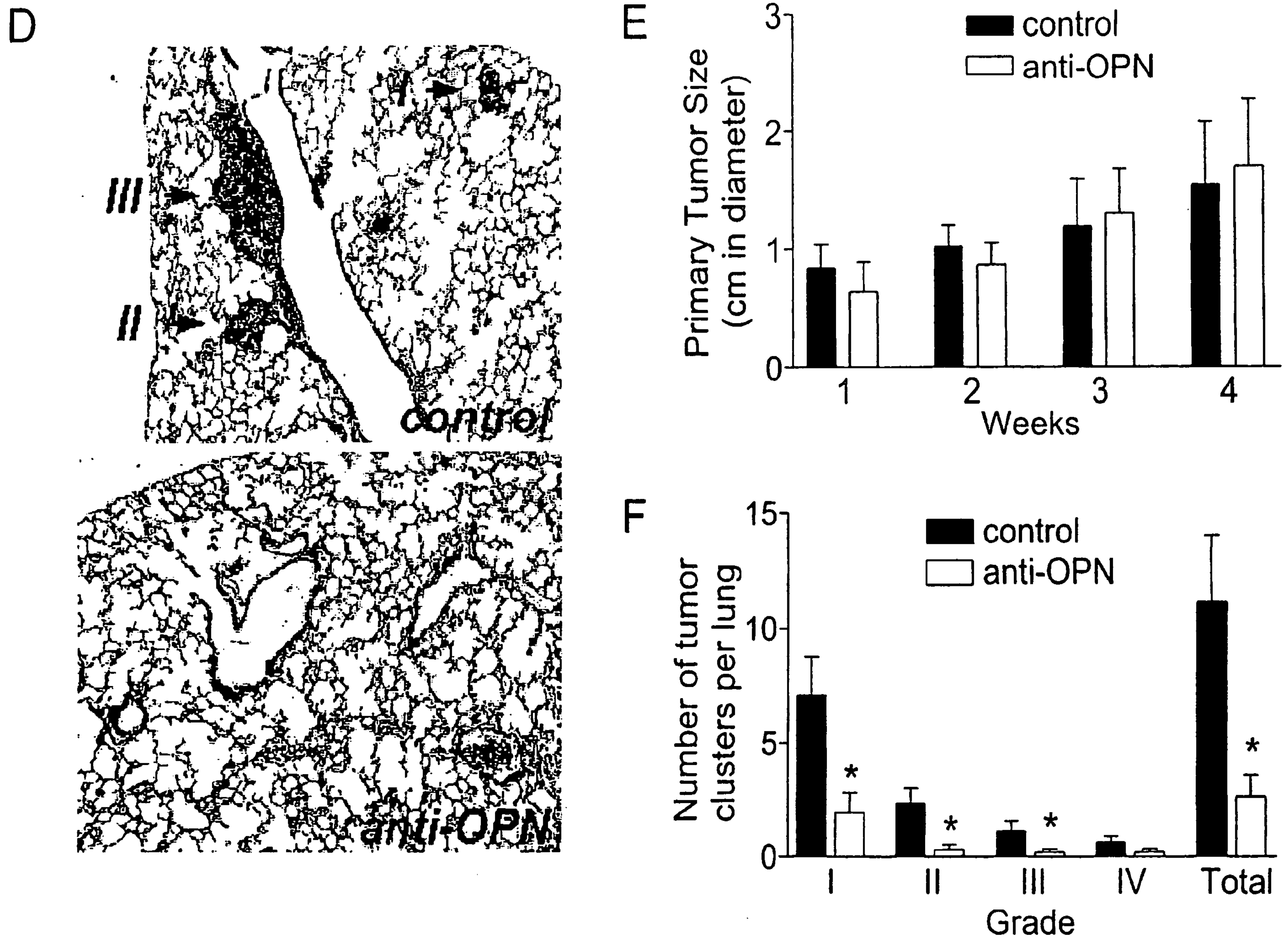
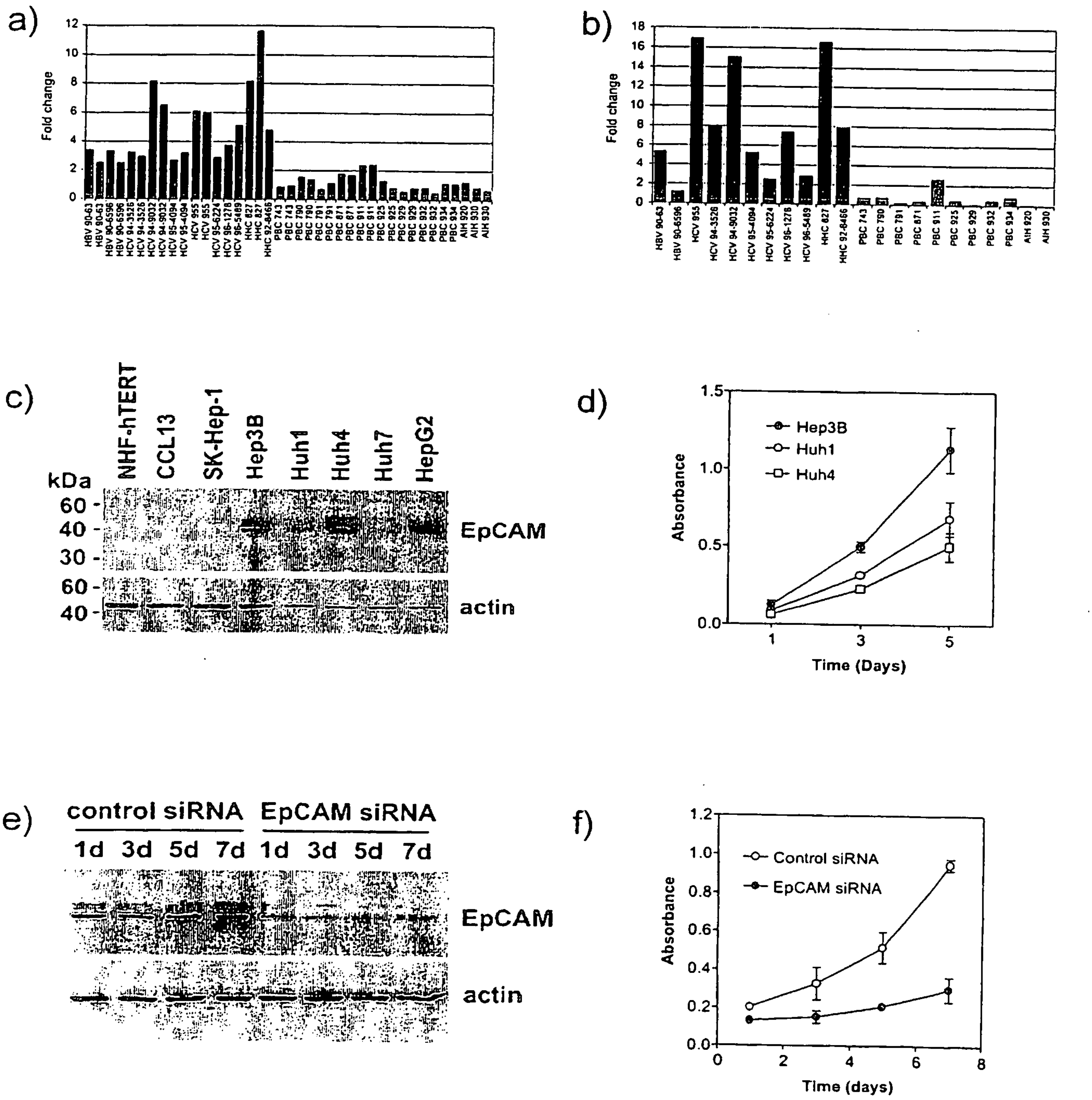

D
III
II
control
anti-OPN
E
Primary Tumor Size (cm in diameter)
0
1
2
3
control
anti-OPN
1
2
3
4
Weeks
F
Number of tumor clusters per lung
0
5
10
15
control
anti-OPN
*
*
*
*
I
II
III
IV
Total
Grade

Fig 6


a)
Fold change
0
2
4
6
8
10
12
HBV 90-63
HBV 90-63
HBV 90-6596
HBV 90-6596
HCV 94-3526
HCV 94-3526
HCV 94-9032
HCV 94-9032
HCV 95-4094
HCV 95-4094
HCV 955
HCV 955
HCV 95-6224
HCV 96-1278
HCV 96-5489
HHC 827
HHC 827
HHC 92-8466
PBC 743
PBC 743
PBC 790
PBC 790
PBC 791
PBC 791
PBC 871
PBC 871
PBC 911
PBC 911
PBC 925
PBC 925
PBC 929
PBC 929
PBC 932
PBC 932
PBC 934
PBC 934
AIH 920
AIH 930
AIH 930
b)
Fold change
0
2
4
6
8
10
12
14
16
18
HBV 90-63
HBV 90-6596
HCV 955
HCV 94-3526
HCV 94-9032
HCV 95-4094
HCV 95-6224
HCV 96-1278
HCV 96-5489
HHC 827
HHC 92-8466
PBC 743
PBC 790
PBC 791
PBC 871
PBC 911
PBC 925
PBC 929
PBC 932
PBC 934
AIH 920
AIH 930
c)
NHF-hTERT
CCL13
SK-Hep-1
Hep3B
Huh1
Huh4
Huh7
HepG2
kDa
60 -
40 -
30 -
EpCAM
60 -
40 -
actin
d)
Hep3B
Huh1
Huh4
Absorbance
0.0
0.5
1.0
1.5
1
3
5
Time (Days)
e)
control siRNA
EpCAM siRNA
1d
3d
5d
7d
1d
3d
5d
7d
EpCAM
actin
f)
Control siRNA
EpCAM siRNA
Absorbance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
Time (days)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 October 2003 (23.10.2003)

PCT

(10) International Publication Number
WO 2003/087766 A3

(51) International Patent Classification⁷: **C12Q 1/68**, C12P 21/06

(21) International Application Number: PCT/US2003/010783

(22) International Filing Date: 4 April 2003 (04.04.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/370,895 5 April 2002 (05.04.2002) US

(71) Applicant *(for all designated States except US)*: **THE GOVERNMENT OF THE UNITED STATES OF AMERICA, as represented by THE SECRETARY OF THE DEPARTMENT OF HEALTH AND HUMAN SERVICES** [US/US]; 6011 Executive Boulevard, Suite 325, Rockville, MD 20852 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **WANG, Xin, Wei** [US/US]; 11409 Crownwood Lane, Rockville, MD 20850 (US). **YE, Qing-Hai** [CN/CN]; Xi Ying Road 33-22-22, Apt. 301, Pu Dong New Area, 200126 Shanghai (CN). **KIM, Jin, Woo** [KR/US]; 12030 Chase Crossing Cir., #404, Rockville, MD 20852 (US).

(74) Agents: WEBER, Kenneth, A. et al.; TOWNSEND AND TOWNSEND AND CREW LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111-3834 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report: 29 July 2004

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2003/087766 A3

(54) Title: METHODS OF DIAGNOSING POTENTIAL FOR METASTASIS OR DEVELOPING HEPATOCELLULAR CARCINOMA AND OF IDENTIFYING THERAPEUTIC TARGETS

(57) Abstract: The present invention relates to methods for diagnosing the metastatic potential of hepatocellular carcinoma (HCC) in HCC patients and methods for diagnosing the potential of developing HCC in patients with chronic liver diseases. A computer readable medium, a digital computer, and a system useful for such diagnosis are also provided. Further disclosed are methods for identifying potential therapeutic targets for treating metastasis in HCC patients and methods for preventing HCC in patients with chronic liver diseases. In addition, the invention provides methods for inhibiting metastasis in HCC patients by suppressing the function of one therapeutic target, osteopontin, and methods for preventing the development of HCC in patients with chronic liver diseases by suppressing the function of one therapeutic target, EpCAM. Pharmaceutical compositions containing agents capable of inhibiting the functions of osteopontin or EpCAM are also disclosed.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/10783

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : C12Q 1/68; C12P 21/06
US CL : 435/6, 69.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 435/6, 69.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,175,084 (INOUE et al) 29 December 1992 (29.12.1992) entire document, especially column 12, lines 34-67, column 13, lines 1-14.	1-27, 34-45
Y	US 6,524,787 B1 (HENDRIX) 25 February 2003, entire document, especially column 3, lines 10-30, column 4, lines 42-67, column 5, lines 1-67, column 6, lines 1-15	1-27, 34-45
Y	US 2003/0211466 A1 (KEENE et al.) 13 November 2003 entire document, especially page 1, paragraph 0008	1-27, 34-45

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
05 April 2004 (05.04.2004)

Date of mailing of the international search report
08 JUN 2004

Name and mailing address of the ISA/US
Mail Stop PCT, Attn: ISA/US
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
Facsimile No. (703) 305-3230

Authorized officer
Valerie Bell-Harris for
Suryaprabha Chunduru
Telephone No. 571/272-1600

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US03/10783

Box I Observations where certain claims were found unsearchable (Continuation of Item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. □ Claim Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. □ Claim Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. □ Claim Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of Item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
Please See Continuation Sheet

1. □ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. □ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. □ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.: 1-27, 34-45

Remark on Protest □ The additional search fees were accompanied by the applicant's protest.
□ No protest accompanied the payment of additional search fees.

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I, claim(s) 1-27, 34-45, drawn to a method for identifying potential therapeutic targets for inhibiting metastasis in a patient suffering from hepatocellular carcinoma.

Group II, claim(s) 28-33, drawn to a computer readable medium and system comprising data sets.

The inventions listed as Groups I-II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: Each of the groups are independent and lack the same special technical feature that links the groups together because the claims in Group are drawn to a method for identifying potential therapeutic targets for inhibiting metastasis in a patient from hepatocellular caricinoma is independent by itself and do not depend on the claims in Group II, which are drawn to a computer readable system.

Continuation of B. FIELDS SEARCHED Item 3:
Biosis, Embase, Medline, LifeSci, Caplus
search terms: hepatocellular carcinoma, diagnosis, gene expression, array